AD________________

Award Number: W81XWH-10-1-0337

TITLE: Identification and Characterization of MYC Regulatory Elements: Links to Prostate Cancer

PRINCIPAL INVESTIGATOR: Nora Wasserman

CONTRACTING ORGANIZATION: The University of Chicago
Chicago, IL 60637

REPORT DATE: September 2012

TYPE OF REPORT: Final Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE November 2012	**2. REPORT TYPE** Final Summary	**3. DATES COVERED** 15 May 2010- 30 August 2012

4. TITLE AND SUBTITLE
Identification and Characterization of MYC Regulatory Elements: Links to Prostate Canter

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-10-1-0337

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Nora Wasserman
Marcelo Nobrega

E-Mail: wasserman@uchicago.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The University of Chicago
Chicago, IL 60637

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Prostate cancer is the most common cancer diagnosed in males in the developed world. Genome-wide association studies (GWAS) have greatly helped in the identification of common risk variants associated with complex diseases such as cancer; routinely, these associated polymorphisms are located within gene deserts and other type of non-coding DNA. A striking example of GWAS implicating non-coding variants in the etiology of cancer can be seen on chromosome 8q24, where numerous studies have reported associations between prostate (and other) cancer and variants concentrated within a 1.2Mb gene desert. Although there are no genes within the interval, the proto-oncogene MYC lies just downstream of the gene desert, raising the possibility that the associated risk regions may harbor long-range cis-regulatory elements – such as enhancers – involved in the tissue-specific transcriptional regulation of MYC. To date, we have located and characterized an in vivo prostate enhancer encompassing the prostate cancer associated SNPrs6983267. Furthermore, we demonstrated that this enhancer exhibits allele-specific activity in developing and mature mouse prostates, mimicking MYC expression. Our findings help advance the field's understanding of the mechanistic reason for the overwhelming association seen between this 8q24 gene desert and prostate cancer.

15. SUBJECT TERMS- none provided

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	51	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

INTRODUCTION:

Prostate cancer is the most common cancer diagnosed in males in the developed world. While risk factors suggest a genetic basis for the disease, the search for causal genes has yielded few results. In the last decade, genome-wide association studies (GWAS) have greatly helped in the identification of common risk variants associated with complex diseases such as cancer; routinely, these associated polymorphisms are located within gene deserts and other type of non-coding DNA (*1*). A striking example of GWAS implicating non-coding variants in the etiology of cancer can be seen on chromosome 8q24, where numerous studies have reported associations between prostate, colorectal, breast and urinary bladder cancer and variants concentrated within a 1.2Mb gene desert (*2-12*). Evidence for prostate cancer association is particularly strong, with five distinct linkage disequilibrium (LD) blocks spanning a 440Kb interval harboring risk variants. Although there are no well-characterized genes within the interval, the proto-oncogene *MYC* lies just downstream of the gene desert, raising the possibility that the associated risk regions may harbor long-range *cis*-regulatory elements – such as enhancers – involved in the tissue-specific transcriptional regulation of *MYC*. Under this hypothesis, each distinct prostate cancer association interval would contain a functional element involved in regulating *MYC* expression in the prostate. The purpose of this proposal is to identify and characterize prostate enhancers within the prostate cancer associated intervals on 8q24 using a combination of *in vivo* and *in vitro* reporter assays.

BODY:

Detailed accounts of progress and results for each task within my Statement of Work can be found below. These achievements support the hypothesis put forth in my initial proposal. They resulted in the characterization of an allele-specific prostate enhancer encompassing the prostate cancer associated SNP rs6983267 and culminated in a publication within a first tier genomics journal (*13*) (see Appendix for full publication).

Task 1a: Perform in situ hybridization for all genes within a 1.0Mb interval surrounding the prostate cancer-associated region.

Because of its status as an ideal positional and functional target gene, we began our *in situ* hybridizations by assessing *Myc* expression in the genitourinary apparatus of male mice. Digoxigenin-labeled *Myc* antisense and sense riboprobes were generated from a full-length mouse *Myc* cDNA clone. Staining was performed on whole P8 and P21 mouse prostates for 48 hours. As expected, we observed *Myc* expression in the developing and mature prostate, as well as in the coagulating glands, seminal vesicles, and ductus deferens (*13*) (Appendix, Figure 2 of paper). This expression correlated very well with the reporter gene expression pattern driven by the rs6983267-containing enhancer we identified (described below).

While the parallel between the enhancer's domain and *Myc* expression in the prostate is compelling, it does not directly show that *MYC* is the target gene for the 8q24 *cis*-regulatory element. In theory, other genes within the enhancer's potential range of influence – including *FAM84B*, *POU5F1P1,* and *PVT1* – could also exhibit prostate expression and be the target for the element's regulatory influence. However, work published since the submission of this proposal has convincingly demonstrated that the 8q24 prostate cancer associated regions (including the rs6983267-containing enhancer we describe) physically interact with the *MYC* promoter in prostate cancer cell lines (*14-16*). These studies all employed chromosomal conformation capture (3C), a technique that assesses whether a specific fragment can loop over large genomic distances to physically connect with another DNA region (*17*). This direct link between the prostate cancer associated regions and the *MYC* promoter –located between 250kb and 650kb away – shows that *MYC* is the target gene of the cancer associated *cis*-regulatory regions. This obviates the need to investigate the expression patterns of other nearby genes.

Task 1b, 1c and 1d: Use progressively smaller DNA fragments in in vivo *mouse transgenic assays to identify and localize enhancers within the 8q24 region.*

To initially examine the 8q24 gene desert for regulatory elements, we surveyed the region using a broad-scale BAC scan approach. We identified three overlapping human BACs encompassing the prostate cancer risk regions (CTD-2506D10, RP11-124F15, and CTD-2533C10), which together span 480kb of non-coding DNA (Appendix, Figure 2 of paper). Each BAC carried the prostate cancer-associated risk haplotype and was tagged through a Tn7 transposon-mediated random insertion of a β-galactosidase (lacZ) gene driven by a β–globin minimal promoter (*18*). The lacZ cassette integration converts the BACs into enhancer trapping systems, whereby any long-range enhancer(s) contained within each ~180kb BAC can act upon the reporter gene to drive tissue- and temporal-specific β-galactosidase expression. The design of overlapping BACs aids in the efficiency of the system to narrow the critical region of interest, as expression profiles unique to only one BAC must be due to uniquely contained sequences; conversely, identical expression patterns present in overlapping BACs suggest that the functional element driving β-galactosidase expression must be contained in the shared genomic region. A detailed account of these experiments can be found in the attached manuscript (Appendix, Results and Methods).

The *in vivo* BAC transgenic reporter assays identified prostate enhancer activity contained within the 8q24 gene desert (Appendix, Figure 1 of paper). While we did not observe β-galactosidase prostate expression in BAC CTD-2506D10 transgenic mice (12 independent transgenics), animals harboring BACs CTD-2533C10 and RP11-124F15 displayed β-galactosidase prostate expression at days P0, P8 and P21 (*13*). Because of the highly similar reporter expression patterns obtained from BACs RP11-124F15 and CTD-2533C10, including prostate, coagulating gland, and urethral/bladder lining, we hypothesized that our BAC transgenic assays were identifying a single prostate enhancer within the 59kb shared genomic segment of these two BACs. Interestingly, one of the most strongly associated prostate cancer risk SNPs, rs6983267, is contained within this 59kb overlapping interval and disrupts an evolutionarily conserved sequence (Appendix, Figure 1 of paper).

Rather than using fosmids as an intermediate means to localize the putative prostate enhancer element, we directly tested the rs6983267-containing evolutionarily conserved element for regulatory potential *in vivo.* A 5kb DNA fragment containing each allele of this SNP was cloned into a lacZ reporter cassette using Invitrogen's Gateway cloning system and transgenic mice harboring either the risk or the non-risk variant of rs6983267 were generated and analyzed. We determined that the conserved sequence containing the prostate cancer GWAS SNP displayed allele-specific *in vivo* prostate enhancer properties (*13*) (Appendix, Figure 2 of paper). Specifically, the risk allele, rs6983267-G, led to consistent, stronger β-galactosidase expression in prostates and coagulating glands than the non-risk allele, rs6983267-T, in P0, P8 and P21 transgenic mice (Appendix, Figures 2 and 3). The expression pattern driven by the rs6983267-G risk allele in 3 independent mouse transgenic lines closely resembled that observed in BACs RP11-124F15 and CTD-2533C10 – both of which also harbor the risk allele. In contrast, the rs6983267-T non-risk allele led to weakened prostate and coagulating gland expression in 3 independent transgenic lines. For each allelic variant evaluated, those transgenic founders exhibiting enhancer activity showed highly concordant β-galactosidase expression in the prostate, with a clear qualitative difference between the risk and non-risk variants.

These results demonstrate that our BAC-based enhancer trapping screen is a powerful resource to rapidly uncover *cis*-regulatory regions in large DNA segments. While we were unable to uncover other prostate-specific regulatory elements within the 8q24 locus, we did identify a mammary gland enhancer within BAC RP11-124F15, a region that has been associated with risk to breast cancer (*13*).

Task 2a: Construct a reporter plasmid for luciferase assay analysis in prostate cancer cell lines.

Two luciferase assay report plasmids were constructed for the quantitative analysis of enhancer potential in prostate cancer cell lines. Both make use of Promega's pGL4 vectors. The first uses the minimum promoter present in the pGL4.23, with the only alteration being the addition of Invitrogen's Gateway cassette into the multiple cloning site. This allows for the easy shuttling of multiple elements into the vector without the need for traditional cloning. The second vector began with Promega's promoterless pGL4.10, into which the *MYC* promoter and the Gateway cassette were both inserted. *MYC* is known to be expressed from numerous promoters, with the majority of transcripts initiating from promoter 2 (P2); as its proximal regulation is still not entirely understood, we wished to be overly conservative in the definition of "promoter" to ensure that all necessary elements were present in our reporter construct (*19*). To that end, 1.7kb of sequence upstream of the *MYC* transcriptional start site was cloned into the pGL4.10 vector.

Task 2b: Optimize transfection conditions for prostate cancer cell lines.

Prostate cancer cell lines DU-145, PC-3, and LNCaP were selected for analysis because of their extensive use throughout the literature. They were cultured according to ATCC recommendations (http://www.atcc.org). Invitrogen's Lipofectamine 2000 was used as a transfection reagent and transfections were performed according to the published protocol. Cell count optimization resulted in plating 100,000 cells/well in a 24 well standard tissue culture plate. Experiments optimizing DNA quantity determined that 1ug of test DNA and 100ng of Renilla plasmid yielded the best transfection results.

Task 2c: Identify risk polymorphisms contained within enhancer elements identified in Aim 1.

Despite our best efforts to characterize other prostate enhancers, we only succeeded in identifying the previously described SNP rs698326-containing enhancer element (*13*). As others have already determined that this GWAS SNP is the causative variant (*20, 21*), this task is irrelevant.

Task 2d: Quantitatively measure the enhancing potential of MYC regulatory elements using both risk and non-risk variants.

Plasmids containing the either the risk or the non-risk haplotype of the 5kb rs698326-containing enhancer element were tested for regulatory potential in three prostate cancer cell lines. Luciferase expression was normalized to the Renilla plasmid and all reactions were performed in triplicate. Although others have shown a measurable difference in luciferase activity in colorectal cancer cell lines (*20, 21*), we were unable to differentiate between the two alleles in our prostate cancer lines. These results directly conflict with the robust findings of our *in vivo* enhancer experiments, and highlight the fact that cell line based assays often have a higher variance based on culture conditions and other artifacts.

Training plan: Skill acquisition and research presentations

During the course of this award, I have successfully mastered and honed the technical skills necessary to complete the body of work outlined above; namely *in situ* hybridizations, BAC recombineering, cloning, animal husbandry and dissection, cell culture techniques, and data analysis/interpretation. In addition, I presented results at regularly schedule lab meetings, gave two departmental talks, and presented a poster at the 2011 *IMPaCT* meeting. I attended many seminars on relevant material, although I did not audit a cancer biology course due to time constraints. In parallel with the culmination of my training grant, I was awarded my Ph.D. in Human Genetics in August, 2012.

KEY RESEARCH ACCOMPLISHMENTS:

- We identified a prostate enhancer located within a prostate cancer associated region capable of driving *in vivo* reporter gene expression in the developing and mature mouse prostate. Furthermore, we showed that the genotype of the cancer associated SNP rs6983267 – contained within this enhancer – conveys allele-specific regulatory potential to the enhancer element, with the risk variant possessing stronger enhancer abilities than the protective allele. These findings were published in the high impact journal Genome Research (*13*).
- Our broad-scale BAC scan of the 8q24 gene desert showed that there are other regulatory sequences within the interval of interest; specifically, we uncovered a mammary gland enhancer within a region that has been associated with risk to breast cancer (*13*) (Appendix, Figure 1 of paper). These results demonstrate that we have generated a powerful tool to experimentally interrogate genomic regions showing association to multiple types of cancer, and that this tool can be widely disseminated among the cancer genetics research community.
- Our results, combined with those of other researchers working with colorectal cancer (*14-16, 22, 23*), demonstrate that the same genetic variation – known to increases risk to both prostate and colorectal cancer – functions in both cases by altering the spatial, temporal, and/or quantitative fine tuning of *MYC* expression through allele-specific enhancer activity.
- As our *in vivo* enhancer reporter assays allow for the interrogation of regulatory potential over developmental time, we were able to demonstrate that the rs6983267-containing enhancer is active throughout prostate organogenesis (*13*) (Appendix, Figure 2 and 3 of paper). These results pose the intriguing possibility that the increased risk to prostate cancer might result from a misregulation of *MYC*'s expression early in development, long before the onset of tumorigenesis.
- We determined that in our hands, cell line-based luciferase assays have difficulty picking up the allele-specific activity of our prostate enhancer. This result highlights the importance of our *in vivo* enhancer-trapping BAC assay as an efficient and valuable tool for discovering and characterizing regulatory elements (*13*).

- Our publication was positively received by the community, and as a result I was asked to write a book chapter discussing *cis*-regulatory mechanisms underlying cancer risk. I did so – focusing on prostate cancer – and the chapter is published in "Gene Regulatory Sequences and Human Disease," edited by Nadav Ahituv, Ph.D (Appendix). To date, this chapter has the most downloads of any in the book and is a valuable resource for the field.

REPORTABLE OUTCOMES:

Publications:

Wasserman NF, Nobrega MA. Cis-Regulatory Variation and Cancer. In: Ahituv, N, ed. *Gene Regulatory Sequences and Human Disease*. 2012 Edition: Springer; 2012:195-216.

Wasserman NF, Aneas I, Nobrega MA. An 8q24 gene desert variant associated with prostate cancer risk confers differential *in vivo* activity to a *MYC* enhancer. *Genome Research* 20(9), 1191-1197 (2010).

Presentations at Scientific Meetings:

Wasserman NF, Nobrega MA. An 8q24 gene desert variant associated with prostate cancer risk confers differential *in vivo* activity to a *MYC* enhancer. Poster. *IMPaCT*, 2011.

Degrees Obtained:

Ph.D. in Human Genetics, University of Chicago. Chicago, IL, August 2012.

CONCLUSION:

The BAC enhancer trapping strategy that we employed allowed us to rapidly interrogate the 440kb of 8q24 prostate cancer-associated non-coding DNA for *cis*-regulatory elements. We effectively screened a half-megabase genomic interval *in vivo* using only three constructs, and succeeded in identifying a prostate enhancer within an interval strongly associated with prostate cancer. In addition, we localized a specific prostate enhancer contained within the overlapping region of two of our BACs and showed that it possessed *in vivo* allele-specific regulatory abilities contingent on the genotype of the prostate cancer associated SNP rs6983267. These results – showing the cancer risk allele demonstrating stronger enhancer potential than the non-risk allele – are concordant with *MYC*'s known role as a proto-oncogene. Finally, we demonstrated that the rs6983267-containing enhancer exhibits differential *in vivo* activity throughout prostate organogenesis. As no association has been seen between rs6983267 genotype and steady-state *MYC* mRNA levels in normal prostate cells or prostate tumors (*24*), our results raise the possibility that this variant asserts its influence on prostate cancer risk before tumorigenesis actually occurs. We also determined that these *in vivo* experiments were more sensitive at highlighting this allele-specific enhancer activity than luciferase assays in a panel of prostate cancer cell lines, emphasizing the importance of *in vivo* studies. Our findings contribute to the field's understanding of the mechanistic reason for the overwhelming association seen between this 8q24 gene desert and prostate cancer. By explaining the genetic basis for disease risk, progress towards clinical applications can be made.

REFERENCES:

1. A. Visel, E. M. Rubin, L. A. Pennacchio, Genomic views of distant-acting enhancers. *Nature* **461**, 199 (Sep 10, 2009).
2. A. A. Al Olama *et al.*, Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet* **41**, 1058 (Oct, 2009).
3. L. T. Amundadottir *et al.*, A common variant associated with prostate cancer in European and African populations. *Nat Genet* **38**, 652 (Jun, 2006).
4. D. F. Easton *et al.*, Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087 (Jun 28, 2007).
5. M. Ghoussaini *et al.*, Multiple loci with different cancer specificities within the 8q24 gene desert. *J Natl Cancer Inst* **100**, 962 (Jul 2, 2008).
6. J. Gudmundsson *et al.*, Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* **39**, 631 (May, 2007).
7. C. A. Haiman *et al.*, A common genetic risk factor for colorectal and prostate cancer. *Nat Genet* **39**, 954 (Aug, 2007).
8. L. A. Kiemeney *et al.*, Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat Genet* **40**, 1307 (Nov, 2008).
9. I. Tomlinson *et al.*, A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* **39**, 984 (Aug, 2007).
10. C. Turnbull *et al.*, Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet* **42**, 504 (Jun, 2010).
11. M. Yeager *et al.*, Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39**, 645 (May, 2007).
12. B. W. Zanke *et al.*, Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* **39**, 989 (Aug, 2007).
13. N. F. Wasserman, I. Aneas, M. A. Nobrega, An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res* **20**, 1191 (Sep, 2010).
14. M. M. Pomerantz *et al.*, The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41**, 882 (Aug, 2009).
15. N. Ahmadiyeh *et al.*, 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc Natl Acad Sci U S A* **107**, 9742 (May 25, 2010).
16. J. Sotelo *et al.*, Long-range enhancers on 8q24 regulate c-Myc. *Proc Natl Acad Sci U S A* **107**, 3001 (Feb 16, 2010).
17. J. Dekker, K. Rippe, M. Dekker, N. Kleckner, Capturing chromosome conformation. *Science* **295**, 1306 (Feb 15, 2002).
18. F. Spitz, F. Gonzalez, D. Duboule, A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113**, 405 (May 2, 2003).
19. I. Wierstra, J. Alves, The c-myc promoter: still MysterY and challenge. *Adv Cancer Res* **99**, 113 (2008).
20. S. Tuupanen *et al.*, The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet.* **41**, 885 (2009).
21. M. M. Pomerantz *et al.*, The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet.* **41**, 882 (2009).
22. S. Tuupanen *et al.*, The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* **41**, 885 (Aug, 2009).
23. J. B. Wright, S. J. Brown, M. D. Cole, Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol Cell Biol* **30**, 1411 (Mar, 2010).
24. M. M. Pomerantz *et al.*, Evaluation of the 8q24 prostate cancer risk locus and MYC expression. *Cancer Res* **69**, 5568 (Jul 1, 2009).

CSH PRESS GENOME RESEARCH

An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a *MYC* enhancer

Nora F. Wasserman, Ivy Aneas and Marcelo A. Nobrega

Genome Res. 2010 20: 1191-1197 originally published online July 13, 2010
Access the most recent version at doi:10.1101/gr.105361.110

An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a *MYC* enhancer

Nora F. Wasserman, Ivy Aneas, and Marcelo A. Nobrega[1]

Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA

Genome-wide association studies (GWAS) routinely identify risk variants in noncoding DNA, as exemplified by reports of multiple single nucleotide polymorphisms (SNPs) associated with prostate cancer in five independent regions in a gene desert on 8q24. Two of these regions also have been associated with breast and colorectal cancer. These findings implicate functional variation within long-range *cis*-regulatory elements in disease etiology. We used an in vivo bacterial artificial chromosome (BAC) enhancer-trapping strategy in mice to scan a half-megabase of the 8q24 gene desert encompassing the prostate cancer-associated regions for long-range *cis*-regulatory elements. These BAC assays identified both prostate and mammary gland enhancer activities within the region. We demonstrate that the 8q24 cancer-associated variant rs6983267 lies within an in vivo prostate enhancer whose expression mimics that of the nearby *MYC* proto-oncogene. Additionally, we show that the cancer risk allele increases prostate enhancer activity in vivo relative to the non-risk allele. This allele-specific enhancer activity is detectable during early prostate development and throughout prostate maturation, raising the possibility that this SNP could assert its influence on prostate cancer risk before tumorigenesis occurs. Our study represents an efficient strategy to build experimentally on GWAS findings with an in vivo method for rapidly scanning large regions of noncoding DNA for functional *cis*-regulatory sequences harboring variation implicated in complex diseases.

[Supplemental material is available online at http://www.genome.org.]

Genome-wide association studies (GWAS) routinely implicate variation within gene deserts and other types of noncoding DNA in the etiology of disease (Houlston et al. 2008; Silverberg et al. 2009; Yang et al. 2009; Liu et al. 2010). A recent meta-analysis of ~1200 disease-associated single nucleotide polymorphisms (SNPs) found that in 40% of cases, known exonic sequences were absent from the associated linkage disequilibrium (LD) blocks (Visel et al. 2009). While the presence of nonannotated transcripts or noncoding RNAs may explain some of the noncoding disease associations, these observations also have been interpreted as evidence that many of the associated noncoding regions harbor variants that alter the activity of long-range *cis*-regulatory elements controlling gene expression. Enhancers are one such type of long-range element, functioning over up to megabase-long genomic distances to regulate the temporal and tissue-specific expression patterns of their target gene(s) (Nobrega et al. 2003). A large number of genes with tissue- and temporal-specific expression patterns are known to be controlled by an array of enhancers, with each individual *cis*-regulatory element driving a subset of its gene's entire expression profile (Carroll 2008). This modular nature of enhancer activity makes them ideal candidates for involvement in complex diseases, as functional variants in an individual *cis*-element would result in changes to gene expression only in specific organs/tissue types.

Despite the plethora of GWAS signals implicating noncoding regions in complex disease risk, strategies to experimentally follow up on such findings are lacking. This deficiency stems principally from the difficulty in identifying functional noncoding sequences that map remotely from their target genes. Programs such as ENCODE have been addressing this deficiency by developing and applying technologies to identify these elusive types of long-range regulatory elements (The ENCODE Project Consortium 2007). While these technologies have been invaluable in the identification of putative functional noncoding sequences, they rely heavily on cell culture and other in vitro and in silico methodology to identify and experimentally validate enhancers and other elements. Thus, although these techniques are ideal for functionally following up on noncoding GWAS results when the relevant cell type of interest is obvious and accessible, problems can arise if the putative element under investigation imparts its transcriptional regulatory effects in a cell type of unpredicted origin or one that is not amenable to routine culture. Necessary, but lagging, is the development of simpler in vivo strategies that can concurrently query the spatial and temporal properties of functional *cis*-regulatory sequences within large segments of noncoding DNA. Our goal in this study is to describe one such strategy for following up on GWAS results, and to test its ability to uncover noncoding risk variants in loci associated with complex diseases.

A striking example of GWAS implicating noncoding variants in the etiology of complex diseases can be seen on chromosome 8q24, where numerous studies have reported associations between multiple types of cancer—including prostate, colorectal, breast, and urinary bladder—and variants concentrated within 620 kb of a 1.2-Mb gene desert (Amundadottir et al. 2006; Easton et al. 2007; Gudmundsson et al. 2007; Haiman et al. 2007; Tomlinson et al. 2007; Zanke et al. 2007; Ghoussaini et al. 2008; Kiemeney et al. 2008; Al Olama et al. 2009). Evidence for prostate cancer association within the region is particularly strong, with five distinct LD blocks spanning a 440-kb interval on 8q24 harboring risk variants (Fig. 1A, all shaded regions; Ghoussaini et al. 2008; Al Olama et al. 2009). One of these prostate cancer-associated variants, rs6983267, is independently associated with colorectal cancer (Fig. 1A, green; Tomlinson et al. 2007), and a second prostate cancer-associated LD

[1] **Corresponding author.**
E-mail nobrega@uchicago.edu.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.105361.110. Freely available online through the *Genome Research* Open Access option.

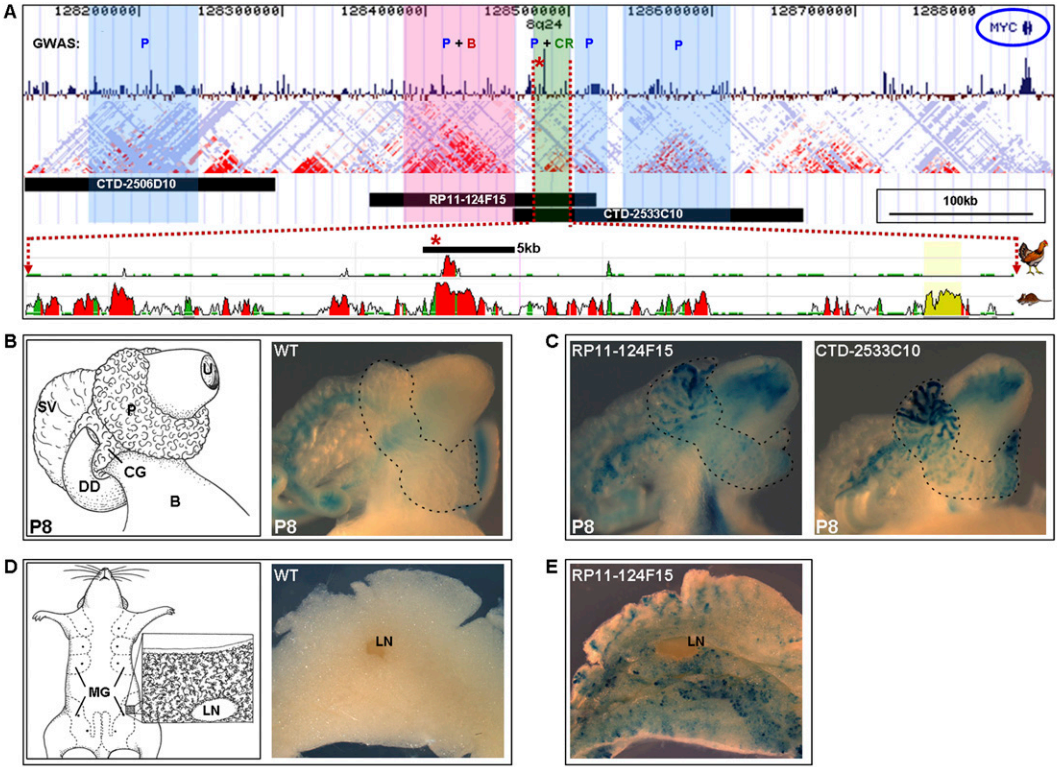

Figure 1. The 8q24 *MYC* gene desert harbors prostate and mammary gland transcriptional enhancers. (*A*) Five susceptibility loci within the 440-kb interval shown to be associated with prostate cancer (all shaded regions; blue denotes a prostate-only association), with one locus independently associated with breast cancer (pink) and a second associated with colorectal cancer (green). (B) Breast cancer–associated region, (CR) colorectal cancer–associated region, (P) prostate cancer–associated region. (Blue circle) *MYC*, (red asterisk) SNP rs6983267. (*Below*) The three human *lacZ*-tagged BACs encompassing the prostate cancer risk regions. (Red dotted lines) The LD block containing SNP rs6983267—associated with both prostate and colorectal cancers and contained within BACs RP11-124F15 and CTD-2533C10—is shown in detail. Sequence conservation is shown in chicken and mouse genomes (human genome used as reference). (*B*) The male genitourinary apparatus in P8 mice, shown as a cartoon (*left*) and in wild-type, nontransgenic mice (*right*). (Dashed line, *right*) Outline of the prostate. (B) Bladder, (CG) coagulating gland, (DD) ductus deferens, (P) prostate, (SV) seminal vesicle, (U) urethra. There is endogenous X-gal staining in the SV and DD. (*C*) Representative P8 prostates from transgenic mice containing BAC RP11-124F15 or CTD-2533C10 showing prostate and urogenital apparatus enhancer activity. (Dashed lines) Outlines of prostates. (*D*) The mammary gland in midgestational pregnant females, shown as a cartoon (*left*) and in wild-type, nontransgenic mice (*right*). The enlargement (*left*) illustrates a lymph node, ducts, and alveoli and in a mammary fat pad. (LN) Lymph node, (MG) mammary gland. (*E*) Representative mammary fat pad from a day 14.5 pregnant female harboring BAC RP11-124F15.

block harbors a distinct SNP (rs13281615) that shows association with breast cancer (Fig. 1A, pink; Easton et al. 2007). Although no well-annotated genes lie within this interval, the independent associated variants (or linked functional elements within the associated regions) may all be regulating the expression patterns of a single gene involved in cancer tumorigenesis and/or progression in various tissue types. The proto-oncogene *MYC* lies immediately downstream of this gene desert, raising the possibility that the associated regions of risk may harbor long-range *cis*-regulatory elements involved in the tissue-specific transcriptional regulation of *MYC* expression; under this hypothesis, each distinct association interval might harbor a functional noncoding element involved in regulating *MYC* expression in the corresponding tissue type for each implicated cancer. A summary of the 8q24 gene desert and its numerous cancer loci is shown in Figure 1. Here, we have chosen to specifically focus on the multiple independent associations between this 8q24 gene desert and prostate cancer.

Encoding a well-known transcription factor essential to the regulation of cell proliferation and growth, *MYC* is up-regulated at both the mRNA and protein levels in aggressive prostate cancers (DeMarzo et al. 2003). In addition, copy-number analyses in prostate cancer specimens have identified the 8q24 region surrounding *MYC* as the most common recurrent region of chromosomal gain (Lapointe et al. 2007). These findings show that prostate cancers employ multiple mechanisms for achieving *MYC* overexpression, through transcriptional up-regulation or through amplification of gene copy number. We hypothesized that variation within *MYC*'s long-range *cis*-regulatory elements could disrupt the quantitative, temporal, or spatial expression patterns of *MYC* in the prostate, possibly underlying the GWAS signals identified in the 8q24 gene desert. In this study, we describe how an in vivo bacterial artificial chromosome (BAC) enhancer-trapping strategy efficiently scanned the 8q24 gene desert for *cis*-regulatory sequences, and report on the identification of both prostate and mammary gland enhancer activities within the assayed regions. We further refined the prostate enhancer interval, showing that it harbors the prostate cancer risk SNP rs6983267, and demonstrate that the two resultant allelic variants display functionally polymorphic prostate enhancer properties in vivo.

Results

Surveying the regulatory landscape of the 8q24 gene desert

To initially examine the 8q24 gene desert for regulatory elements, we surveyed the region using a broad-scale BAC scan approach

(Spitz et al. 2003). This strategy allows for the rapid and effective examination of large genomic regions for *cis*-regulatory elements, and can be readily applied to any locus of interest. We identified three overlapping human BACs encompassing the prostate cancer risk regions (Fig. 1A), which together span 480 kb of noncoding DNA. Each BAC carried the prostate cancer-associated risk haplotype and was tagged through a Tn7 transposon-mediated random insertion of a beta-galactosidase (*lacZ*) gene driven by a beta-globin minimal promoter (Spitz et al. 2003). The transposon-mediated insertion was performed using simple, commercially available kits (see Methods) and occurs in vitro; the protocol yields rapid results and can be easily scaled up for the simultaneous tagging of numerous BACs.

The *lacZ* cassette integration converts the BACs into enhancer-trapping systems, whereby any long-range enhancer(s) contained within each ~180-kb BAC can act upon the reporter gene to drive tissue- and temporal-specific beta-galactosidase expression. Any enhancers present within a given BAC are then simultaneously interrogated using a reporter assay system, allowing for the concurrent examination of large genomic regions for functional noncoding elements. The design of overlapping BACs aids in the efficiency of the system to narrow the critical region of interest, as expression profiles unique to only one BAC must be due to uniquely contained sequences; conversely, identical expression patterns present in overlapping BACs suggest that the functional element driving beta-galactosidase expression must be contained in the shared genomic region. Modified BACs were analyzed by PCR and pulsed-field gel electrophoresis to confirm the integration of the Tn7β-*lacZ* reporter cassette. To mitigate any possible effects of unknown insulator or silencer elements within the BAC sequence, we selected clones with at least two Tn7β-*lacZ* integration events. Each BAC was then injected into fertilized mouse oocytes to generate transgenic mice in accordance with IACUC regulatory standards. For each BAC, a minimum of two independent transgenic founders were obtained and studied; this is necessary to overcome potential position-dependent expression effects resulting from random integration of the transgene (BAC).

We assayed *lacZ* expression at multiple points in prostate organogenesis and maturation; postnatal days 0 and 8 (P0 and P8) during prostate development, and P21, when prostate maturation is virtually complete (Sugimura et al. 1986). At each developmental stage, prostates were dissected and stained for beta-galactosidase expression using X-gal (Fig. 1B,C; Kothary et al. 1989).

These in vivo BAC transgenic reporter assays identified prostate enhancer activity contained within the 8q24 gene desert (Fig. 1C). While we did not observe beta-galactosidase prostate expression in BAC CTD-2506D10 transgenic mice (12 independent transgenics), animals harboring BACs CTD-2533C10 and RP11-124F15 displayed beta-galactosidase prostate expression at days P0 (data not shown), P8 (Fig. 1C), and P21 (data not shown). As illustrated in Figure 1C, the beta-galactosidase expression domain of both BAC RP11-124F15 and BAC CTD-2533C10 extends to other components of the urogenital system, including the coagulating glands, urethra, and the lining of the urinary bladder. While the seminal vesicles and ductus deferens also exhibit X-gal staining, we and others observed this expression pattern in both wild-type (Fig. 1B) and transgenic animals, reflecting the presence of endogenous beta-galactosidase in these structures (Wang et al. 2002; Krajnc-Franken et al. 2004). As 80% of the prostatic ducts are formed by day P15 in mice (Sugimura et al. 1986), our data indicate that the enhancer(s) contained within these two BACs are active both during and after prostate organogenesis and maturation.

Because some of the prostate cancer-associated regions also have been associated with breast and colorectal cancer (Fig. 1A), we chose to additionally assay the mammary glands, colon, and rectum of those animals transgenic for BACs containing the relevant regions (BAC RP11-124F15 for breast cancer, and both BACs RP11-124F15 and CTD-2533C10 for colorectal cancer). Mammary glands were examined at embryonic day 14.5 (E14.5), when the mammary buds have fully formed in female embryos, in 11-wk-old virgin females with mature branched glands, and in prelactating females 14 d after conception, when the mammary gland undergoes extensive hyperplasia and tissue remodeling (Hens and Wysolmerski 2005; Oakes et al. 2006; Sternlicht 2006).

We observed in vivo mammary gland enhancer activity in mice transgenic for BAC RP11-124F15 (Fig. 1E), which harbors associated intervals for not only prostate but also breast and colorectal cancer. Transgenic animals displayed beta-galactosidase expression in the epithelial compartment—ducts and alveoli (Hennighausen and Robinson 2005)—of the mammary glands of midgestational pregnant and 11-wk-old virgin females (Fig. 1E; data not shown). No enhancer activity was seen in E14.5 embryos. Of note, Jia et al. (2009) recently identified a noncoding element within this region capable of in vitro enhancer activity in breast cancer cell lines; this element should be viewed as a strong candidate for the mammary gland activity we see in vivo.

Characterizing the prostate enhancer

We next aimed to refine the location of the prostate enhancer(s) within the BACs driving prostate expression. Because of the highly similar reporter expression patterns obtained from BACs RP11-124F15 and CTD-2533C10, including prostate, coagulating gland, and urethral/bladder lining, we hypothesized that our BAC transgenic assays were identifying a single prostate enhancer within the 59-kb shared genomic segment of these two BACs. Interestingly, one of the most strongly associated prostate cancer risk SNPs, rs6983267, is contained within this 59-kb overlapping interval and disrupts an evolutionarily conserved sequence (Fig. 1A).

To directly test the rs6983267-containing evolutionarily conserved element for regulatory potential in vivo, we cloned a 5-kb DNA fragment containing each allele of this SNP in a *lacZ* reporter cassette using Invitrogen's Gateway cloning system (Kothary et al. 1989). Transgenic mice harboring either the risk or the non-risk variant of rs6983267 were generated and analyzed. We determined that the conserved sequence containing the prostate cancer GWAS SNP displayed allele-specific in vivo prostate enhancer properties (Fig. 2). Specifically, the risk allele, rs6983267-G, led to consistent, stronger beta-galactosidase expression in prostates and coagulating glands than the non-risk allele, rs6983267-T, in P0, P8, and P21 transgenic mice (Figs. 2A,B, 3B,C). The expression pattern driven by the rs6983267-G risk allele in three independent mouse transgenic lines closely resembled that observed in BACs RP11-124F15 and CTD-2533C10—both of which also harbor the risk allele. In contrast, the rs6983267-T non-risk allele led to weakened prostate and coagulating gland expression in three independent transgenic lines (Fig. 2B). For each allelic variant evaluated, those transgenic founders exhibiting enhancer activity showed highly concordant beta-galactosidase expression in the prostate, with a clear qualitative difference between the risk and non-risk variants.

To test whether this spatial reporter expression pattern of the rs6983267-containing enhancer correlates with endogenous *MYC* expression in prostate and other components of the urogenital

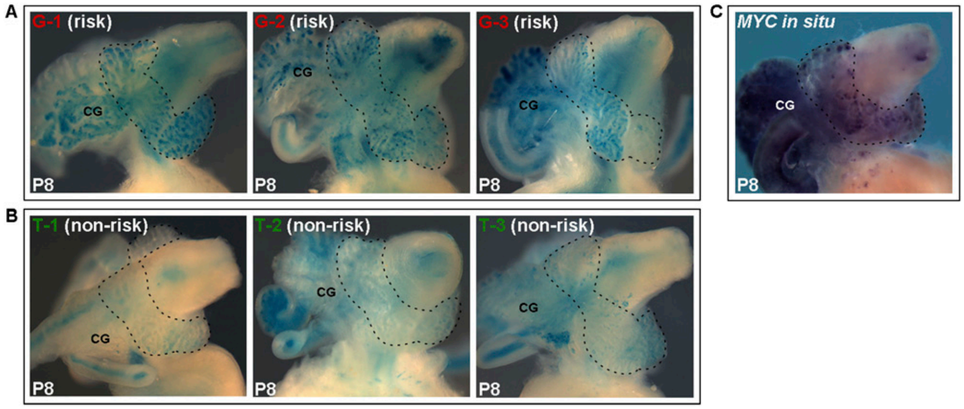

Figure 2. SNP rs6983267 mediates allelic-specific enhancer activity in mouse prostates. Three independent transgenic founders harboring reporter plasmids driven by either the G (risk) allele (*A*) or T (non-risk) allele (*B*) are shown at P8. (Dashed lines) Outlines of prostates; (CG) coagulating glands. The prostate cancer risk allele leads to consistently stronger beta-galactosidase expression in prostates and coagulating glands than the non-risk allele in vivo. (*C*) *MYC* in situ hybridization at P8 correlates with the reporter expression pattern driven by the rs6983267-containing enhancer.

system, we performed whole mount in situ hybridizations using a full-length *Myc* probe in mouse prostates at P8 (Wilkinson and Nieto 1993). We observed *Myc* expression in the male genitourinary apparatus, including the prostate, in a pattern closely mimicking the reporter expression of the rs6983267-G enhancer and BACs CTD-2533C10 and RP11-124F15, both of which harbor the G risk allele as well (Fig. 2C).

This same prostate enhancer that we have characterized also has been shown to act as an allelic-specific long-range *MYC* enhancer in colorectal cancer cells (Jia et al. 2009; Pomerantz et al. 2009a; Tuupanen et al. 2009; Wright et al. 2010). Although we did not observe colorectal enhancer activity in our initial BAC screen of the region, we again assayed transgenic animals harboring either the risk or non-risk rs6983267-containing enhancer element for in vivo enhancer activity in the colorectal area at three developmental time points. We observed no beta-galactosidase expression in E14.5 intestines for either construct tested, and colorectal X-gal staining at P8 and P21 was indistinguishable between wild-type mice and transgenic animals harboring either enhancer variant (Supplemental material). Strong endogenous beta-galactosidase expression is observed in intestines of both wild-type and transgenic animals starting at E15.5, limiting our ability to identify in vivo colorectal enhancers in late embryogenesis and postnatally. These findings highlight the difficulty in assaying postnatal in vivo intestinal enhancers using *lacZ* reporter assays.

Investigations into the embryonic activity of the rs6983267-containing element demonstrated that while this enhancer has several spatial domains of expression, its allele-specific activity is restricted to the prostate and coagulating glands. Both the rs6983267-G and rs6983267-T enhancer elements drove expression in several spatial domains of E11.5 and E14.5 embryos, with no apparent allelic-specific enhancer activity (Fig. 3A). Transgenics harboring either haplotype variant showed similar X-gal staining in the limbs and tail at E11.5, consistent with previously reported patterns (data not shown; Tuupanen et al. 2009). We also observed enhancer activity in the developing urinary bladder, genital tubercle, and limbs in the E14.5 embryos. This pattern, which precedes prostate development, is also indistinguishable between the allelic variants of this enhancer (Fig. 3A).

Taken together, our data posit that the rs6983267-containing enhancer is part of *MYC*'s regulatory landscape, and that the variant within this enhancer may increase the risk of prostate cancer through its role in allelic-specific control of *MYC* expression in the prostate.

Discussion

The BAC enhancer-trapping strategy that we employed allowed us to rapidly interrogate the 440 kb of 8q24 prostate cancer-associated noncoding DNA for *cis*-regulatory elements. We effectively screened a half-megabase genomic interval in vivo using only three constructs, identifying the existence of mammary gland and prostate enhancers in the interval associated with each respective

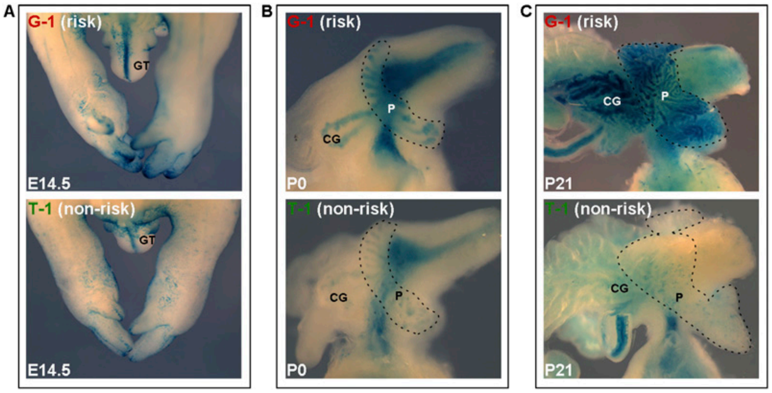

Figure 3. The rs6983267-containing enhancer demonstrates distinct temporal regulatory abilities. Representative G (risk, *top*) and T (non-risk, *bottom*) transgenics are shown at a series of developmental time points. (*A*) E14.5 transgenic embryos exhibit beta-galactosidase expression in the genital tubercle and limbs, with no apparent allele-specific enhancer activity. (GT) Genital tubercle. (*B*,*C*) Allele-specific regulatory ability is visible in neonatal P0 pups (*B*) and P21 adolescent mice (*C*), with in vivo prostate and coagulating gland beta-galactosidase expression qualitatively stronger in the risk allele (*top*) line than the non-risk variant (*bottom*). (CG) Coagulating gland, (P) prostate.

cancer type. We believe that this methodology provides a significant advance to current genomic techniques for following up on GWAS results in noncoding regions, as it can be easily adapted to examine loci in vivo on a megabase scale. As demonstrated by our results, this strategy can be used to concurrently identify spatially and temporally unique enhancers within a large sequence, and can be useful in refining the critical regions for enhancer mapping, while still permitting the use of a whole-systems, in vivo animal model.

These relatively straightforward BAC transgenic reporter assays also provide a way to more closely approximate the genomic context of relevant enhancers. By testing ~200 kb of sequence simultaneously, enhancers are assayed in a context much closer to their true genomic environment, one where they are subjected to (largely unknown) modifications by neighboring repressors, insulators, chromatin changes, and/or various other interactions with nearby *cis* sequences. In traditional plasmid-based reporter assays, this important genomic context is lost. We conducted our clone selection strategy so as to minimize the potential negative effects of such insulators or repressors; tagged BACs containing at least two copies of the Tn7β-*lacZ* reporter cassette—integrated near each end of the BAC sequence—were selected for experimental use. We hypothesized that this would diminish false-negative results caused by repressive elements in a single-copy integration clone. When compared with BACs tagged with just a single Tn7β-*lacZ* cassette, we observed more reproducible results in mice transgenic for BACs harboring two Tn7β-*lacZ* integrations (M.A.N, unpubl.).

Because we observed the same urogenital system spatial pattern of expression in both of the overlapping BACs tested, we deduced that the enhancer was within the small interval shared between those BACs. However, it is possible that other prostate enhancers also exist within the BACs we tested. To formally exclude this possibility, other approaches could have been used, including the analysis of additional enhancer-trapping BACs with complementary overlapping patterns. Alternatively, BAC recombineering could have been employed to specifically delete our known enhancer from the BACs assayed. Both approaches are logical follow-ups to the in vivo BAC transgenic reporter assays, and would maintain the analytical strengths of assaying enhancers in their genomic environments.

Recent studies have reported on the colorectal and prostate enhancer activities of the rs6983267-containing sequence we describe here (Jia et al. 2009; Pomerantz et al. 2009a; Tuupanen et al. 2009; Sotelo et al. 2010; Wright et al. 2010). Using a combination of genome-wide in vitro assays, this sequence has been highlighted as possessing attributes of an enhancer, including specific chromatin modifications and binding of transcription factors. Several groups have demonstrated that in colorectal cancer cell lines, TCF7l2 (TCF4) binds preferentially to the risk allele (rs6983267-G) of this enhancer (Pomerantz et al. 2009a; Tuupanen et al. 2009; Wright et al. 2010). Reports regarding the enhancer properties of this sequence in prostate cancer cell lines have been mixed, however. When tested in LNCaP and PC3 prostate cancer cell lines, this sequence displayed enhancer properties only in the former, possibly due to the PC-3 line's lack of androgen receptor expression (Jia et al. 2009). In a second study, this rs6983267-containing enhancer was unable to drive luciferase expression above promoter-only levels in LNCaP or PC-3 cells, unless cells were cotransfected with Tcf4 and beta-catenin expression vectors (Sotelo et al. 2010). Under those conditions, the rs6983267-containing element demonstrated allelic-specific enhancer activity in LNCaP cells, but with the non-risk rs6983267-T variant driving stronger expression than the risk rs6983267-G allele.

Our in vivo results—showing the cancer risk allele demonstrating stronger enhancer potential than the non-risk allele—corroborate those reported in colorectal cancer cell lines (Pomerantz et al. 2009a; Tuupanen et al. 2009; Wright et al. 2010), and are concordant with *MYC*'s known role as a proto-oncogene. Our whole-animal experimental strategy obviated the experimental variation added by cell lines to clearly show that this element is a functional prostate enhancer in vivo, while also adding the ability to investigate enhancer activity throughout organogenesis. We believe that this broad spatial and temporal characterization of regulatory potential is ideally afforded by in vivo experimentation, and propose this as the standard in the follow-up to GWAS risk variants implicated in human disease.

The rs6983267-containing element physically interacts with *MYC*'s promoter in both colorectal cancer and prostate cancer cell lines, providing evidence that this enhancer is involved in regulating *MYC* expression in these two tissue types (Pomerantz et al. 2009a; Sotelo et al. 2010; Wright et al. 2010). Despite these compelling findings and the fact that altered *MYC* expression has been implicated repeatedly in the pathogenesis of prostate cancers (Williams et al. 2005), no association has been seen between rs6983267 genotype and *MYC* mRNA levels in normal prostate cells or prostate tumors (Pomerantz et al. 2009b). This lack of genotype–phenotype correlation implies that steady-state *MYC* mRNA levels in adult prostate tissue may not be the correct biological entity underlying risk. Our findings demonstrate that the rs6983267-containing enhancer exhibits differential in vivo activity throughout prostate organogenesis, and raise the possibility that this variant asserts its influence on prostate cancer risk long before tumorigenesis occurs. With widely varying risk allele frequencies in different populations—from 49% in American Caucasians to 81% in African Americans (HapMap, merged Phase 1, 2, and 3 frequencies)—this SNP may also have an effect on the population prevalence of both prostate cancer and colorectal cancer (Jemal et al. 2009).

We have described how a noncoding SNP strongly associated with disease can in fact alter the in vivo activity of its encompassing *cis*-regulatory element, suggesting a possible impact on cancer risk before tumorigenesis actually occurs. Although further studies are warranted, our in vivo temporal data hint at an underlying molecular explanation for this nongenic SNP's contribution to prostate cancer risk. These findings emphasize the notion that thorough investigations into the regulatory impact of polymorphisms are an indispensable component to the functional follow-up of GWAS scans, and stress the importance of conducting these experiments using in vivo systems.

Methods

Transposon-mediated BAC modification

BACs CTD-2506D10, RP11-124F15, and CTD-2533C10 were modified by in vitro random transposition of Tn7β-*lacZ* (Spitz et al. 2003). BAC DNA was extracted by using the Nucleobond AX Kit (Macherey-Nagel). Twenty nanograms of Tn7β-*lacZ* vector was mixed with 20–40 ng of BAC DNA, GPS buffer, and TnsABC transposase (New England BioLabs), followed by incubation for 10 min at 37°C. Start solution was added and the reaction was extended for 1 h. After heat inactivation for 10 min at 75°C and a 1-h dialysis, electrocompetent DH10B cells were transformed with 2 μL of the transposition reaction. Cells were plated on LB agar containing

20 μg/mL kanamycin and 20 μg/mL chloramphenicol. Positive colonies were first identified by polymerase chain reaction (PCR) using beta-globin and *lacZ* primers (Tn7β-*lacZ* beta-globin F: AGCA TCTATTGCTTACATTTGC; Tn7β-*lacZ lacZ* R: ATAGGTTACGTTGG TGTAGATGG). Modified BAC clones were then digested with NotI and separated by pulsed-field gel electrophoresis overnight on a 1% agarose gel to determine the number of copies and the position(s) of the integrated Tn7β-*lacZ* cassette. Clones with two copies of the cassette were chosen for further analysis to minimize the possible influence of silencer or insulator elements with the BACs.

lacZ plasmid generation

The 5 kb of sequence surrounding the rs6983267-containing conserved element was PCR amplified from human genomic DNA heterozygous for the rs6983267 SNP (rs6983267 F: TCTTGACCTG ATTGCTGAAAAAT; rs6983267 R: TCTGGGGGTGAGTTAAATGA TAA). The fragment was then purified using the QIAquick PCR Purification Kit (Qiagen) and cloned into the pDONR 221 Gateway entry vector (Invitrogen). Colonies were analyzed by restriction enzyme analysis for successful fragment insertion, and positive clones were sequenced to determine the allelic status of SNP rs6983269 (rs6983267-seq F: TAGACACCAAGAGGGAGGTATCA; rs6983267-seq R: CCAGGTTAAAGGAAACTGAACTG). Clones containing sequence harboring both the risk (G) and non-risk (T) rs6983267 allele were transferred to a Gateway-HSP68-*lacZ* reporter vector using the LR recombination reaction (Invitrogen) (Poulin et al. 2005). All plasmids were again verified by restriction analysis and direct sequencing prior to pronuclear mouse injections.

Production of transgenic mice

Tn7β-*lacZ* tagged BAC DNA was purified using the Nucleobond BAC 100 Kit (Macherey-Nagel), rehydrated in injection buffer (10 mM Tris at pH 7.5; 0.1 mM EDTA), and diluted to a concentration of 2 ng/μL. BAC DNA was injected in its circular form.

Plasmid DNA was purified using the Plasmid Maxi Kit (Qiagen), and 50 μg of each plasmid was digested with SalI to excise the vector backbone. Following a gel purification step using the QIAquick Gel Extraction Kit (Qiagen), the DNA to be injected was further purified using a standard ethanol precipitation. The purified DNA was dialyzed for 24 h against injection buffer (10 mM Tris at pH 7.5; 0.1 mM EDTA), and its concentration was determined fluorometrically and by agarose gel electrophoresis. The DNA was diluted to a concentration of 2 ng/μL. Purified BAC and plasmid DNA were then used for pronuclear injections of CD1 mouse embryos in accordance with standard protocols approved by the University of Chicago.

For the Tn7β-*lacZ* tagged BACs, multiple stable transgenic lines were generated for each construct, and F_1 animals were analyzed for each line at multiple postnatal developmental time points. BAC CTD-2506D10 DNA injections yielded 12 independent lines (0/12 positive for prostate beta-galactosidase expression); injections of RP11-124F15 and CTD-2533C10 both resulted in two independent beta-galactosidase-expressing lines.

For the rs6983267-containing enhancer plasmid, a total of three beta-galactosidase-expressing independent transgenics was obtained for rs6983267-G; three beta-galactosidase-expressing independent transgenic animals/lines were also obtained for rs6983267-T. For several of these independent lines, the F_0 animals themselves were analyzed at P8; this excluded any analysis of the line at other time points. For the risk allele, rs6983267-G, we obtained two F_0 animals positive for beta-galactosidase expression in the prostate. The third independent rs6983267-G transgenic was maintained as a stable line. For the non-risk allele, rs6983267-T, one F_0 transgenic animal was obtained; the remaining two independent transgenics were maintained as stable lines.

Mouse in vivo transgenic reporter assay

Prostates and mammary glands were harvested from mice at P0, P8, and P21 and dissected into cold 100 mM phosphate buffer (PBS) (pH 7.3), followed by 30–45 min of incubation with 4% paraformaldehyde at 4°C. E14.5 embryos were incubated in 4% paraformaldehyde for 2 h. Tissues were then washed two times for 20 min with wash buffer (2 mM $MgCl_2$; 0.01% deoxycholate; 0.02% NP-40; 100 mM phosphate buffer at pH 7.3), and stained for 18 h at room temperature with freshly made staining solution (0.8 mg/mL X-gal; 4 mM potassium ferrocyanide; 4 mM potassium ferricyanide; 20 mM Tris at pH 7.5 in wash buffer). After staining, samples were rinsed five times for 20 min in PBS and post-fixed in 4% paraformaldehyde. For each animal analyzed, tail samples were taken at the time of dissection and DNA was isolated through the addition of lysis buffer (100 mM Tris-HCl at pH 8.5, 5 mM EDTA, 0.2% SDS, 200 mM NaCl, and 1 mg/mL proteinase K) and incubation overnight at 55°C. Genotyping was performed by PCR with primers within the reporter cassette/vector (using beta-globin and *lacZ* primers for the Tn7β-*lacZ* tagged BACs, rs6983267-seq primers for the plasmids).

Imaging

All photographs were taken using a Leica MZ16 F stereomicroscope and QCapture Pro software. Settings (lighting, exposure time) were kept constant between structure- and aged-matched samples. Images displayed in the paper were generated using an image processing software package (CombineZM) that allows for the creation of extended depth of field images. Multiple pictures of each structure were taken at varying depth of fields and then computationally integrated; the focus areas are blended to create a composite high-resolution image with an extended depth of field. This allowed for the production of images where all the multiple plains of the urogenital apparatus appear well focused and defined.

In situ hybridization

In situ hybridization analysis on whole P8 prostates using digoxigenin-labeled *Myc* antisense and sense riboprobes was performed according to standard protocols (Wilkinson and Nieto 1993). The probes were generated from a full-length mouse *Myc* cDNA clone (IMAGE ID 3962047). Staining was performed for 48 h, and the stained prostates were then transferred to 10% buffered formalin phosphate prior to imaging.

Acknowledgments

We thank François Spitz for kindly providing us with the beta-globin-Tn7 vector and Linda Degenstein for assistance in generating transgenic animals. We also thank James Noonan, Rick Kittles, and Gail Prins for their consultation and support. The urogenital apparatus and mammary gland cartoons in Figure 1, B and D, were kindly drawn by John Westlund. This work was partially supported by grant HG004428 (M.A.N.). N.F.W. is supported by a DoD Prostate Cancer Training Award (PC094251).

References

Al Olama AA, Kote-Jarai Z, Giles GG, Guy M, Morrison J, Severi G, Leongamornlert DA, Tymrakiewicz M, Jhavar S, Saunders E, et al. 2009. Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet* **41:** 1058–1060.

Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, Agnarsson BA, Sigurdsson A, Benediktsdottir KR, Cazier JB, Sainz J, et al. 2006. A common variant associated with prostate cancer in European and African populations. *Nat Genet* **38:** 652–658.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* **134:** 25–36.

DeMarzo AM, Nelson WG, Isaacs WB, Epstein JI. 2003. Pathological and molecular aspects of prostate cancer. *Lancet* **361:** 955–964.

Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, et al. 2007. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447:** 1087–1093.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Ghoussaini M, Song H, Koessler T, Al Olama AA, Kote-Jarai Z, Driver KE, Pooley KA, Ramus SJ, Kjaer SK, Hogdall E, et al. 2008. Multiple loci with different cancer specificities within the 8q24 gene desert. *J Natl Cancer Inst* **100:** 962–966.

Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, et al. 2007. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* **39:** 631–637.

Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, Waliszewska A, Neubauer J, Tandon A, Schirmer C, McDonald GJ, et al. 2007. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet* **39:** 638–644.

Hennighausen L, Robinson GW. 2005. Information networks in the mammary gland. *Nat Rev Mol Cell Biol* **6:** 715–725.

Hens JR, Wysolmerski JJ. 2005. Key stages of mammary gland development: Molecular mechanisms involved in the formation of the embryonic mammary gland. *Breast Cancer Res* **7:** 220–224.

Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S, Chandler I, Vijayakrishnan J, Sullivan K, Penegar S, et al. 2008. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* **40:** 1426–1435.

Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. 2009. Cancer statistics, 2009. *CA Cancer J Clin* **59:** 225–249.

Jia L, Landan G, Pomerantz M, Jaschek R, Herman P, Reich D, Yan C, Khalid O, Kantoff P, Oh W, et al. 2009. Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet* **5:** e1000597. doi: 10.1371/journal.pgen.1000597.

Kiemeney LA, Thorlacius S, Sulem P, Geller F, Aben KK, Stacey SN, Gudmundsson J, Jakobsdottir M, Bergthorsson JT, Sigurdsson A, et al. 2008. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat Genet* **40:** 1307–1312.

Kothary R, Clapoff S, Darling S, Perry MD, Moran LA, Rossant J. 1989. Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. *Development* **105:** 707–714.

Krajnc-Franken MA, van Disseldorp AJ, Koenders JE, Mosselman S, van Duin M, Gossen JA. 2004. Impaired nipple development and parturition in LGR7 knockout mice. *Mol Cell Biol* **24:** 687–696.

Lapointe J, Li C, Giacomini CP, Salari K, Huang S, Wang P, Ferrari M, Hernandez-Boussard T, Brooks JD, Pollack JR. 2007. Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. *Cancer Res* **67:** 8504–8510.

Liu Y, Blackwood DH, Caesar S, de Geus EJ, Farmer A, Ferreira MA, Ferrier IN, Fraser C, Gordon-Smith K, Green EK, et al. 2010. Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. *Mol Psychiatry* doi: 10.1038/mp.2009.107.

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302:** 413.

Oakes SR, Hilton HN, Ormandy CJ. 2006. The alveolar switch: Coordinating the proliferative cues and cell fate decisions that drive the formation of lobuloalveoli from ductal epithelium. *Breast Cancer Res* **8:** 207. doi: 10.1186/bcr1411.

Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA, Chan JA, Hills A, Davis M, et al. 2009a. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41:** 882–884.

Pomerantz MM, Beckwith CA, Regan MM, Wyman SK, Petrovics G, Chen Y, Hawksworth DJ, Schumacher FR, Mucci L, Penney KL, et al. 2009b. Evaluation of the 8q24 prostate cancer risk locus and MYC expression. *Cancer Res* **69:** 5568–5574.

Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, Pennacchio LA. 2005. In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics* **85:** 774–781.

Silverberg MS, Cho JH, Rioux JD, McGovern DP, Wu J, Annese V, Achkar JP, Goyette P, Scott R, Xu W, et al. 2009. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat Genet* **41:** 216–220.

Sotelo J, Esposito D, Duhagon MA, Banfield K, Mehalko J, Liao H, Stephens RM, Harris TJ, Munroe DJ, Wu X. 2010. Long-range enhancers on 8q24 regulate c-Myc. *Proc Natl Acad Sci* **107:** 3001–3005.

Spitz F, Gonzalez F, Duboule D. 2003. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113:** 405–417.

Sternlicht MD. 2006. Key stages in mammary gland development: The cues that regulate ductal branching morphogenesis. *Breast Cancer Res* **8:** 201. doi: 10.1186/bcr1368.

Sugimura Y, Cunha GR, Donjacour AA. 1986. Morphogenesis of ductal networks in the mouse prostate. *Biol Reprod* **34:** 961–971.

Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M, Wood W, et al. 2007. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* **39:** 984–988.

Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Bjorklund M, Wei G, Yan J, Niittymaki I, et al. 2009. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* **41:** 885–890.

Visel A, Rubin EM, Pennacchio LA. 2009. Genomic views of distant-acting enhancers. *Nature* **461:** 199–205.

Wang Y, Newton DC, Miller TL, Teichert AM, Phillips MJ, Davidoff MS, Marsden PA. 2002. An alternative promoter of the human neuronal nitric oxide synthase gene is expressed specifically in Leydig cells. *Am J Pathol* **160:** 369–380.

Wilkinson DG, Nieto MA. 1993. Detection of messenger RNA by in situ hybridization to tissue sections and whole mounts. *Methods Enzymol* **225:** 361–373.

Williams K, Fernandez S, Stien X, Ishii K, Love HD, Lau YF, Roberts RL, Hayward SW. 2005. Unopposed c-MYC expression in benign prostatic epithelium causes a cancer phenotype. *Prostate* **63:** 369–384.

Wright JB, Brown SJ, Cole MD. 2010. Upregulation of c-MYC in *cis* through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol Cell Biol* **30:** 1411–1420.

Yang JJ, Cheng C, Yang W, Pei D, Cao X, Fan Y, Pounds SB, Neale G, Trevino LR, French D, et al. 2009. Genome-wide interrogation of germline genetic variation associated with treatment response in childhood acute lymphoblastic leukemia. *JAMA* **301:** 393–403.

Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E, et al. 2007. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* **39:** 989–994.

Received January 15, 2010; accepted in revised form June 3, 2010.

5. *Cis*-Regulatory Variation and Cancer

Nora F. Wasserman and Marcelo A. Nobrega
Dept. of Human Genetics, University of Chicago

Abstract

In the traditional model of human disease genetics, mutations in coding regions of the genome were assumed to underlie disease phenotypes. It is only in the recent past that functional non-coding regions – such as promoters, enhancers and silencers – have been implicated in disease states. At its most basic level, cancer is a disease caused by the misexpression of genes normally responsible for regulating cell proliferation. It is therefore logical that mutations and variants within *cis*-regulatory elements controlling the expression of proto-oncogenes and tumor suppressor genes would underlie some tumorigenic gene expression changes. As changes in non-coding functional elements are harder to identify than alternations in protein coding sequences, many of the recent insights into *cis*-regulatory variants involved in cancer etiology have been uncovered by genome wide association studies (GWAS) highlighting risk variants in non-genic regions. Here, we highlight examples of cancer-associated variation in promoters, enhancers and silencers, as well as changes to the overall architecture of a gene's regulatory landscape. These functional characterizations bring us closer to understanding the role of *cis*-regulatory mutations and cancer risk/progression.

5.1 Introduction

Cancer is the uncontrolled proliferation of abnormal cells in the body. At the most basic level, this uncontrolled growth is caused by the misexpression of genes normally responsible for regulating cell division. In a healthy cell, the cell cycle is a tightly controlled process, with numerous checkpoints in place to ensure genomic integrity and functioning cell cycle machinery before allowing a cell to proceed into the next phase of the cycle. If DNA damage (caused either by random replication errors or environmental mutagens) is found, the process of division is either paused to allow time for repair or, if the damage is too great, the cell undergoes apoptosis. When proto-oncogenes – genes that positively regulate proliferation or negatively regulate apoptosis – are overexpressed, or tumor suppressor genes – those that negatively control the cell cycle or promote apoptosis – are underexpressed, the cellular checkpoints necessary for controlled division may be less rigorously executed or bypassed entirely. If the burden of mutations impacting the expression of oncogenes and tumor suppressor genes becomes great enough, uncontrolled proliferation can occur and a potentially cancerous cell is created.

The genetic reasons underlying the misexpression of proto-oncogenes and tumor suppressor genes can vary greatly. For proto-oncogenes to become oncogenes, mutations must result in an overexpression of gene product or expanded expression domain (improper spatial or temporal gene activation). This overexpression can be achieved through an increase in gene copy number – where entire chromosomes or chromosomal segments are duplicated or localized genic regions are highly amplified – or through mutations in *cis*-regulatory elements involved in the control of gene expression. These *cis*-regulatory elements include promoters and long-range enhancer or repressor elements that function to regulate gene expression in a tissue- and temporal-specific manner. Enhancing mutations or variations within positive regulatory elements (promoters or enhancers) or weakening alterations to negative regulatory elements (repressors) can result in increased gene expression. Variation within or misuse of enhancer and repressor elements can also contribute to the phenomenon of expanded oncogene expression domain; mutations in enhancers could cause them to take on new functional roles and translocations can result in an enhancer element inappropriately activating a gene near the chromosomal breakpoint. Another mechanistic way for proto-oncogenes to morph into oncogenes is when modifications to protein structure (mutations or deletions) cause them to become constitutively active.

In the inverse scenario, mutations resulting in a decreased level of gene product are necessary for the oncogenic misexpression of tumor suppressor genes. In order for gene expression to be completely silenced, both copies of a tumor suppressor gene must be inactivated. This can be accomplished through any combination of two genetic changes that cause the complete ablation of gene product from one allele, such as the deletion of a gene or entire chromosomal region, a point mutation or frame shift that yields a null allele, or the hypermethylation of a promoter that silences expression. Some tumor suppressor genes also exert oncogenic effects on a cell when their expression levels are simply reduced, rather than eliminated. This can be the result of haploinsufficiency – where expression is totally lost from just one allele – or it can be caused by an overall decrease in the amount of transcription from one or both alleles. In the case of decreased expression from a locus, *cis*-regulatory variation in the promoter or long-range enhancer/repressor elements controlling gene expression is often responsible.

In this chapter, we will focus specifically on *cis*-regulatory mutations and common variation underlying cancer etiology or risk. As touched on above, these *cis*-regulatory underpinnings to gene misexpression represent just a small subset of known genetic alterations involved in the complexities of cancer biology. In many cases, the same genes have been identified as misexpressed in pre-cancerous or cancerous cells due to a multitude of different mechanisms: a particular tumor suppressor gene that is present in a region frequently deleted in tumors may also be the target of an enhancer element containing common variation that exhibits differential activity in a relevant tissue type. This phenomenon highlights the idea that genes critical to controlling cell proliferation will be focus-points for oncogenic mutations, and those mutations may take on many different forms. Many of the more recently discovered examples of *cis*-regulatory changes underlying cancer seem to result in relatively small changes in gene expression levels due to common genetic variation, and therefore have relatively small effect sizes. Because of this, most have been discovered in the functional follow-up to GWAS. The case studies presented here will illustrate instances where *cis*-regulatory changes in promoter, enhancer, and repressor elements that function to modify gene expression levels have been implicated in the etiology of cancer risk.

5.2 Promoter Variation

Located directly upstream of their target gene, promoter elements are the easiest of *cis*-regulatory elements to identify. As the central element involved in controlling gene transcription, their importance and regulatory code has been understood for much longer than long-range *cis*-elements such as enhancers and repressors. As such, countless promoter mutations have been characterized, each altering the expression of a tumor suppressor or proto-oncogene involved in every conceivable type of cancer. Many of these changes – while recurrent in key oncogenic genes – are point mutations unique to a particular individual’s tumor.

As a whole they have taught much about tumor biology, but their invidivual *cis*-regulatory mechanisms of misexpression are not necessarily applicable to wide range of patients. It has only been with the relatively recent advance of GWAS that common variants influencing the regulatory ability of promoters have been identified. Here, we discuss two examples of such GWAS-identified promoter variants, while acknowledging that these represent the very tip of the promoter mutation iceberg.

5.2.1 MSMB and prostate cancer risk:

The most straightforwardly interpreted cases of GWAS hits occur when a potentially functional SNP within an ideal functional candidate gene is found to be associated with a disease. Such was the case when two independent GWAS reported an associated between SNP rs10993994 on 10q11 and prostate cancer risk (Eeles et al. 2008; Thomas et al. 2008). The SNP is 57 base pairs upstream of the transcriptional start site (TSS) of microseminoprotein beta (*MSMB*), a member of the Ig binding factor family known to be a biomarker for prostate cancer and a suggested prostate cancer tumor suppressor gene (Beke et al. 2007; Reeves et al. 2006). Furthermore, rs10993994 had previously been shown to affect promoter activity levels in embryonic kidney cells (Buckland et al. 2005).

Based on this appealing context, two groups set out fine map the associated linkage disequilibrium (LD) block with the goal of showing that the common variation in the *MSMB* promoter was the underlying reason for the prostate cancer association (Chang et al. 2009; Lou et al. 2009). Using independent populations, both groups determined that the GWAS SNP rs10993994 was most strongly associated with prostate cancer risk. To determine the functional significance of this variant, the *MSMB* promoter region – harboring either the risk (T) or the protective (C) allele of rs10993994 – was cloned into a luciferase vector and the promoter activity levels were evaluated in prostate cancer cell lines. Chang et al found that the promoter element containing the T risk allele drove luciferase expression at 13% compared to the protective C allele in LNCaP prostate cancer cells (Chang et al. 2009); this directionality of affect was expected due to *MSMB*'s status as a tumor suppressor gene. The T risk allele also had decreased promoter activity in PC3 prostate cancer cells, as well as in 293T and MCF7 cells lines (Lou et al. 2009).

Once the allele-specific *cis*-regulatory ability of rs10993994 was determined, the question became how the variant exerted its affect on *MSMB* transcriptional activity. As the SNP disrupts a predicted CREB binding site, Lou et al performed electrophoretic mobility shift assays (EMSA) on nuclear extracts of a prostate cancer cell line to see whether the differential CREB binding dependent on the haplotype (Lou et al. 2009). They showed that CREB bound strongly to the protective T allele of rs10993994, where as CREB binding was undetectable in the risk allele. This suggests that the prostate cancer risk SNP modulates *MSMB* promoter activity through differential CREB binding (Lou et al. 2009). Strengthening the evidence for rs10993994's role in *MSMB* expression, Lou et al also showed that cancer cell lines with at least one C allele showed a higher mean *MSMB* mRNA level compared to TT homozygotes (Lou et al. 2009).

To further the link between *MSMB* and prostate cancer tumorigenesis, Pomerantz et al built on the functional studies and investigated the relationship between rs10993994 and *MSMB* expression in normal prostate and prostate tumor samples (Pomerantz et al. 2010). They determined that rs10993994 genotype correlates with *MSMB* mRNA levels in normal and cancerous human prostate cancer specimens, but not in normal colon or breast tissue. This suggests that rs10993994 shows allele-specific activity in a tissue-specific manner. Furthermore, the authors demonstrated that suppression of *MSMB* in prostate epithelial cells resulted in a significant increase in anchorage-independent colony growth; this affect was not seen in mammary epithelial cells (Pomerantz et al. 2010). Taken together, these results show that the *MSMB* promoter SNP rs10993994 exhibits allele-specific *cis*-regulatory activity, and that its affect on *MSMB* expression appears to be prostate specific, in concordance with its status as a common prostate cancer risk variant.

5.2.2 FOXE1 and thyroid cancer risk:

Another example of a promoter *cis*-regulatory variant identified through association studies is the *FOXE1* variant on chromosome 9q22 that was linked to thyroid cancer risk. First identified in a GWAS (Gudmundsson et al. 2009), variants in *FOXE1* were independently flagged as associated with thyroid cancer in a candidate gene association study (Landa et al. 2009). An ideal candidate gene for mis-regulation in thyroid cancer; *FOXE1* is at the center of the regulatory network that initiates thyroid differentiation, and increases in FOXE1 expression correlate with dedifferentiation in thyroid carcinomas (Parlato et al. 2004; Sequeira et al. 2001).

Once the thyroid cancer associated LD block harboring *FOXE1* was located, Landa et al set about assessing all variants within the interval to prioritize candidate causative SNPs. Bioinformatic analysis identified SNP rs1867277 – located 283 bases upstream of the *FOXE1* TSS – as disrupting predicted transcription factor binding sites (TFBS); this variant therefore became the lead candidate for functional analysis. In EMSAs performed with the rs1867277 risk or protective allele and nuclear extracts from a thyroid cancer cell line, a lower band was seen forming with both alleles, while an upper band was found only with the A (risk) allele (Landa et al. 2009). After evaluating predicted TFBS, the authors determined that a Kv channel interacting protein 3, calsenilin (KCNIP3; DREAM) antibody supershifted lower EMSA band complex, while a upstream transcription factor (USF) antibody supershifted the A-specific upper band. They therefore concluded that only the risk A allele of SNP rs1867277 is able to bind transcription factors USF1/USF2. While DREAM overexpression has been previously associated with thyroid enlargement (Rivas et al. 2009), an oncogenic role for the ubiquitously expressed USF1/USF2 factors in thyroid cancer has not yet been established. To further understand the role played by DREAM and the USF1/2 transcription factors in *FOXE1* regulation, luciferase reporter constructs containing one of the two *FOXE1* promoter haplotypes were co-transfected into HeLa cells with cDNA plasmids for DREAM or USF1/2 (Landa et al. 2009). While the DREAM co-transfection did not generate variations in promoter activity, co-transfection of the *FOXE1* promoter with USF1/2 yielded an 8-fold increase in luciferase expression with the A risk allele, but no change with the G protective variant. This data suggests that the differential binding of USF1/2 to the *cis*-regulatory promoter SNP rs1867277 modulates *FOXE1* expression, explaining the region's association with thyroid cancer risk.

5.3 Common variation in long-range *cis*-regulatory elements

Located up to a megabase away from their target gene (Nobrega et al. 2003), long-range *cis*-regulatory elements – such as enhancers and silencers – are functional non-coding elements responsible for controlling tissue- and temporal-specific gene expression. Many key developmental genes are known to be controlled by an array of enhancers, with each individual *cis*-regulatory element driving a subset of its gene's entire expression profile. This modular nature makes them ideal candidates for involvement in complex diseases – like cancer – especially, as a functional variant in an individual *cis*-element would result in changes to gene expression levels only in specific organs/tissue types. Less well-characterized are negative *cis*-

regulatory elements impacting gene expression; although fewer examples exist, they too are presumed to contain functional variation underlying complex disease etiology. As GWAS routinely implicate variation within gene deserts and other types of non-coding DNA in the cancer risk, strategies have been developed for identifying and then characterizing long-range *cis*-regulatory elements potentially harboring cancer-associated variants. The following case studies illustrate examples of successful or in-progress attempts to definitively link non-coding variation with cancer risk.

5.3.1 *MYC* and the 8q24 gene desert cancer associations

The best characterized example of *cis*-regulatory variation in long-range enhancer elements underlying cancer risk was found in chromosome 8q24. Numerous GWAS reported associations between multiple types of cancer – including prostate, colorectal, breast, urinary bladder, and chronic lymphocytic leukemia – and variants concentrated within 620kb of a 1.2Mb gene desert in this region (Al Olama et al. 2009; Amundadottir et al. 2006; Crowther-Swanepoel et al. 2010; Easton et al. 2007; Ghoussaini et al. 2008; Gudmundsson et al. 2007; Haiman et al. 2007b; Kiemeney et al. 2008; Tomlinson et al. 2007; Turnbull et al. 2010; Yeager et al. 2007; Zanke et al. 2007). Thus far, 14 independent polymorphisms have been associated with various cancers in this region (Grisanzio and Freedman 2010), suggesting that multiple independent functional elements underlie disease risk. Although there are no well-annotated genes within the associated intervals, the independent risk variants (or linked functional elements within the associated regions) may all be involved in regulating the expression pattern of a single gene involved in cancer tumorigenesis and/or progression in various tissue types. The infamous proto-oncogene v-myc myelocytomatosis viral oncogene homolog (*MYC*) lies immediately downstream of this gene desert, raising the possibility that the associated regions of risk harbor long-range *cis*-regulatory elements involved in the tissue-specific transcriptional regulation of *MYC* expression; under this hypothesis, each distinct association interval would harbor a functional non-coding element involved in regulating *MYC* expression in the corresponding tissue type for each implicated cancer. Encoding a well-known transcription factor essential to the regulation of cell proliferation and growth, *MYC* is upregulated at both the mRNA and protein level in each of the 8q24 associated cancers (Chen and Olopade 2008; DeMarzo et al. 2003; Nesbit et al. 1999). Additionally, 8q24 is one of the most common regions for somatic amplification in cancer (Beroukhim et al. 2010). *MYC* misregulation due to variation within *cis*-regulatory elements would provide yet another path to its oncogenic overexpression.

In the years following the publication of these striking GWAS results, numerous groups using several complimentary methods have shown that the cancer-associated 8q24 risk regions do in fact harbor enhancer elements (Ahmadiyeh et al. 2010; Jia et al. 2009; Pomerantz et al. 2009; Sotelo et al. 2010; Tuupanen et al. 2009; Wasserman et al. 2010; Wright et al. 2010). The most compelling work centers around the cancer risk variant rs6983267, which has independently been associated with prostate and colorectal cancer (Haiman et al. 2007a; Tomlinson et al. 2007; Yeager et al. 2007; Zanke et al. 2007). SNP rs6983267 is not only the actual typed GWAS variant, but it also disrupts an evolutionarily conserved sequence; this makes it an ideal candidate for functionality. Resequencing and thorough analysis of LD in the cancer-associated region also suggested that rs6983267 itself was the causal risk variant (Yeager et al. 2008). Based on these findings, Pomerantz et al performed targeted chromatin immunoprecipation (ChIP) assays on the evolutionary conserved sequence containing rs6983267 with antibodies known to annotate enhancer elements (Pomerantz et al. 2009). These specific epigenetic marks (such as the histone modification H3K4me1) and proteins (like the coactivator p300) have been shown to reliably mark regulatory regions (Heintzman et al. 2007; Visel et al. 2009). Pomerantz et al found that in the colorectal cancer cell line tested, the rs6983267 element exhibited the classic chromatin signatures for enhancer activity; these findings have since been replicated independently by other groups in both colorectal and prostate cancer cell lines (Ahmadiyeh et al. 2010; Jia et al. 2009; Wright et al. 2010).

While chromatin marks are suggestive of enhancer activity, the regulatory potential of a DNA fragment must be directly assessed using reporter assays. Such experiments ask whether a candidate element is capable of turning on the expression of a reporter gene – usually luciferase for cell-based assays or β-galactosidase for *in vivo* experimentation – in the presence of a minimal promoter. The rs6983267-containing element has been shown to exhibit enhancer activity in colorectal (Jia et al. 2009; Pomerantz et al. 2009; Sotelo et al. 2010; Tuupanen et al. 2009) and prostate (Jia et al. 2009; Sotelo et al. 2010) cancer cell lines, as well as in the developing and mature prostate of transgenic mice (Wasserman et al. 2010). Although cell line-based assays are incredibly useful and relevant to the study of misexpression in cancer cells, the full spatial and temporal characterization of an element's endogenous regulatory potential is ideally afforded by *in vivo* experimentation. It is therefore of particular relevance that the rs6983267-containing enhancer is capable of driving reporter gene expression in the mouse prostate.

If SNP rs6983267 is a *cis*-regulatory modifier of cancer risk, the two alleles would be expected to differentially affect enhancer potential. This allele-specific enhancer activity has in fact been documented in colorectal cancer cell lines (Pomerantz et al. 2009; Tuupanen et al. 2009; Wright et al. 2010) and mouse prostates (Wasserman et al. 2010). In all four cases, the G risk allele was shown to exhibit stronger enhancer activity than the T protective allele in the cancer-relevant cell type. Of note in the *in vivo* system is the fact that the allele-specific enhancer potential seemed to be spatially restricted to the prostate and urogenital apparatus; enhancer activity in the genital tubercle and limbs of E14.5 embryos did not exhibit differential activity between the G and T alleles. Given this enhancer's connection to the proto-oncogene *MYC* (detailed below) in prostate and colorectal cancer, the presumed upregulation in the relevant tissue type caused by the presence of the risk variant fits with the model of misexpression needed for oncogenic change.

Once the regulatory potential of the rs6983267-containing element and the allele-specific nature of the SNP itself was determined, the question as to the mechanistic reason for the differential activity was addressed. The cancer risk variant lies within a predicted TCF consensus binding sequence (Pomerantz et al. 2009; Tuupanen et al. 2009). TCF7L2 is a transcription factor in the Wnt signaling pathway – which is known to target *MYC* – and is activated in most colorectal cancers (Bienz and Clevers 2000; He et al. 1998). Not only was TCF7L2 shown to bind to the rs6983267-containing element in colorectal cancer cell lines, but Pomerantz et al and Tuupanen et al both demonstrated allele-specific binding abilities corresponding to the two rs6983267 alleles: TCF7L2 has a higher affinity for the G risk allele and preferentially binds to that haplotype in heterozygous cells (Pomerantz et al. 2009; Tuupanen et al. 2009). It has also been shown that TCF7L2 binds to the rs6983267-containing element in a prostate cancer cell line (Sotelo et al. 2010). These results suggest that the cancer associated variant mediates risk through differential binding of TCF7L2 to the enhancer element.

The body of work described above convincingly shows that colorectal and prostate cancer associated SNP rs6983267 is located within an enhancer element and that the SNP confers allele-specific activity to its enhancer through (at least in part) the differential binding of

TCF7L2. It does not, however, provide any link – other than circumstantial chromosomal location – between the *cis*-regulatory element and its target gene. In order to definitively associate the enhancer with *MYC*, the ideal candidate gene for misregulation underlying cancer risk, the long-range regulatory element must be shown to physically interact with *MYC*'s promoter. This can be done through the use of the chromosomal conformation capture (3C) assay, a technique that assesses whether a specific fragment (in this case the rs6983267-containing element) can loop over large genomic distances to physically connect with another DNA region (such as the *MYC* promoter, approximately 335kb away) (Dekker et al. 2002). Numerous groups have now demonstrated that the long-range *cis*-regulatory element of interest does in fact interact with *MYC*'s promoter in both colorectal cancer and prostate cancer cell lines, providing very compelling evidence that the rs6983267-containing enhancer is functionally involved in regulating levels of *MYC* expression in these two tissue types (Ahmadiyeh et al. 2010; Pomerantz et al. 2009; Sotelo et al. 2010; Wright et al. 2010). These results provide a crucial link between the *cis*-regulatory risk variant and an infamous proto-oncogene known to be misregulated in the two relevant cancers.

While none of the other 8q24 gene desert risk loci have been as definitively functionally characterized as the LD block harboring the rs6983267-containing element, there is strong evidence for the existence of other long-range tissue-specific *MYC* enhancers within the cancer-associated region boundaries. Two groups have used chromatin marks to identify candidate regulatory elements located in the different association intervals for cell line-based reporter assay tests, and both reported that several exhibited regulatory potential in the relevant cancer cell line (Jia et al. 2009; Sotelo et al. 2010). *In vivo* data also exists for a mammary gland enhancer element contained within the breast cancer LD block, but the precise location of the *cis*-

regulatory element has not yet been determined (Wasserman et al. 2010). Ahmadiyeh et al provided additional support for the hypothesis of multiple *MYC* enhancers throughout the 8q24 gene desert by demonstrating that the cancer associated risk loci physically interact with the *MYC* promoter in a cell type-specific manner. Their 3C results show that breast cancer locus (but not the prostate or colorectal cancer loci) loops to interact with *MYC* in a breast cancer cell line, and that the multiple prostate cancer loci (but not the breast or colorectal cancer loci) physically interact with *MYC* in a prostate cancer cell line (Ahmadiyeh et al. 2010). Taken together, these observations suggest that each distinct cancer association interval does indeed harbor a functional *cis*-regulatory element involved in modulating *MYC* expression in the corresponding tissue type for each implicated cancer. As has been proven for the rs6983267-containing element, the hypothesis remains that each of the *MYC* enhancers harbor variation that influences *MYC* misregulation and cancer risk.

5.3.2. *FGFR2* and breast cancer risk

Another example of *cis*-regulatory variation underlying cancer phenotypes can be seen in the relationship between an intronic region of fibroblast growth factor receptor 2 (*FGFR2)* and breast cancer risk. SNPs within this non-coding LD block exhibited the strongest associations with breast cancer susceptibility in two independent GWAS (Easton et al. 2007; Hunter et al. 2007). Substantiating the strong GWAS association, *FGFR2* – a known breast cancer oncogene – harbors activating missense mutations in some tumors and is somatically amplified in others (Katoh 2008); this makes it an ideal candidate for an additional *cis*-regulatory driven mechanism of misexpression in breast cancer patients.

Meyer et al began their inquiries in the locus by determining that *FGFR2* is expressed at higher levels in breast cancer tumors homozygous for the intronic risk alleles than in tumors homozygous for the protective variants (Meyer et al. 2008). They took this correlation as evidence for a *cis*-regulatory variant within the cancer associated region and focused on identifying differential transcription factor binding abilities for the eight most strongly associated SNPs. EMSA showed that two of the eight candidate functional SNPs (rs7895676 and rs2981578) displayed an allele-specific binding pattern when assayed with nuclear extracts from a breast cancer cell line. By performing supershift experiments, the authors determined that the protective allele of SNP rs7895676 was binding the CCAAT/enhancer binding protein, beta (C/EBPβ), with the risk allele showing no binding affinity. In the case of SNP rs2981578, only the risk allele was capable of binding the runt-related transcription factor 2 (Runx2) (Meyer et al. 2008). Both C/EBPβ and Runx2 have been previously implicated in breast cancer etiology: C/EBPβ is highly overexpressed in malignant breast cells (Grigoriadis et al. 2006) and increased Runx2 expression in breast cancer tumors is associated with a worse clinical outcome (Onodera et al. 1111).

While informative for determining whether DNA-protein complexes are able to form with a given sequence, EMSA cannot establish whether such interactions actually occur within cells. To determine whether the breast cancer risk SNP sites were occupied by the transcription factors of interest in the cellular context, ChIP experiments in breast cancer cell lines homozygous for either the risk or protective haplotype were performed (Meyer et al. 2008). Meyers et al showed differential binding of Runx2 to SNP rs2981578, with the risk allele binding twice as much protein. For rs7895676, the protective allele was enriched for C/EBP; these results support the EMSA findings. The two variants of both SNPs were tested then for allele-specific regulatory ability in breast cancer cell line luciferase reporter assays. The risk allele of rs2981578 stimulated expression when compared to the protective allele protective, while rs7895676 showed weaker results in the opposite direction (with the protective allele displaying stronger potential) (Meyer et al. 2008). When the two SNPs were tested together in one haplotype construct – similar to *in vivo* conditions – the Runx2 SNP prevailed and the risk haplotype showed increased expression. The authors therefore concluded that SNP rs2981578 is likely the functional SNP, as this directionality correlates with increased *FGFR2* expression in tumors harboring risk alleles.

A second study on the same *FGFR2* breast cancer association was performed by Udler et al, using complimentary methods that strengthen the *cis*-regulatory conclusions reached in the previously described work (Udler et al. 2009). Taking advantage of the different haplotype structure present in populations of African decent, the authors fine-mapped the cancer associated region in African American women and concluded that SNP rs2981578 is most strongly associated with breast cancer risk. They also investigated the chromatin state of the region of interest, reasoning that functional *cis*-regulatory elements must be accessible to transcription factors in order to effectively influence target gene expression. DNase I hypersensitivity assays performed in breast cancer cell lines showed that only two SNPs mapped to open chromatin: rs2981578 was one of them (Udler et al. 2009). As it is also within a region of sequence conservation, they concluded that it's likely to be the functional SNP that is influencing breast cancer risk. Taken together, these two studies provide compelling evidence that SNP rs2981578 lies within an active enhancer element and differentially controls its regulatory potential through allele-specific Runx2 binding. While neither of these studies physically links the rs2981578-containing enhancer element to *FGFR2*, *FGFR2* expression in tumors does correlate with SNP genotype and it is an ideal functional candidate for *cis*-regulatory oncogenic misregulation in breast cancer.

5.3.3. *SMAD7* and colorectal cancer risk

The two previous cases illustrated examples where presumed upregulation of oncogenes due to over-active enhancer elements modulated disease risk. This story represents the inverse case, where a cancer risk variant decreases the enhancer activity of an apparent tumor suppressor gene. Several GWAS identified colorectal cancer risk variants on 18q21 within a 17kb LD block in *SMAD7* (Broderick et al. 2007; Curtin et al. 2009; Tenesa et al. 2008), an intracellular antagonist of TGF-beta signaling known to influence colorectal cancer progression (Levy and Hill 2006; ten Dijke and Hill 2004). The associated interval spans both exonic and non-coding sequence, but resequencing excluded coding variations (Broderick et al. 2007).

Lower *SMAD7* expression has been shown to be associated with 18q21 risk variants in lymphoblastoid cell lines (LCLs) (Broderick et al. 2007); assuming that the causal variant was therefore asserting its risk affect through *cis*-regulatory means. Pittman et al. resequenced the entire colorectal cancer associated LD block in a panel of individuals with the goal of identifying all possible variation influencing *SMAD7* expression in the colon (Pittman et al. 2009). The strongest association with disease was provided by a novel SNP dubbed "Novel 1" (rs58920878), which is conserved down to mouse. *In vivo Xenopus* reporter assays performed to determine whether the region surrounding SNP Novel 1 possessed regulatory potential showed GFP expression in the muscle and colorectum of transgenic tadpoles; this strongly suggests that the Novel 1-containing element has enhancer activity (Pittman et al. 2009). Furthermore, the authors demonstrated that the variant confers allele-specific enhancer activity, with the risk allele driving weaker reporter gene expression in the gut compared to the protective haplotype. EMSA results using nuclear extracts from a colorectal cancer cell line revealed the protective allele forming stronger DNA-protein complexes relative to the risk allele, confirming the differential nature of the two alleles (Pittman et al. 2009). The identity of the differentially-bound protein remains unknown, and no definitive link has been established between this enhancer element and the presumed target gene *SMAD7*.

5.3.4. EIF3H and colorectal cancer risk

While enhancers and repressors both fall into the category of long-range *cis*-regulatory elements, much more is known about (and many more examples exist of) enhancers. This is largely due to the existence of more developed methodology for identifying and functionally characterizing these positive regulators. One example of variation within a negative regulatory element can be seen in the functional follow-up to several GWAS that identified risk variants for colorectal cancer on 8q23 within a 300kb region (Houlston et al. 2008; Middeldorp et al. 2009; Tomlinson et al. 2008). After generating a fine-scale map of the region, Pittman et al determined that a 22kb block of LD – located 140kb away from the nearest gene *EIF3H* – showed the highest association with disease (Pittman et al. 2010). Following a similar methodology to the previously described case, they resequenced the associated region in a panel of individuals and prioritized four of the most strongly associated fine-mapped SNPs (rs16892766, “Novel 28,” rs16888589, rs11986063) based on their location within (or flanking) three evolutionally conserved elements. These three conserved elements and their internal/flanking associated SNPs were cloned and tested for *in vivo* enhancer activity in *Xenopus*, zebrafish, and mouse reporter gene transgenic assays. To the authors’ surprise, none of the elements exhibited enhancer activity (Pittman et al. 2010). Luciferase reporter assays in colorectal cancer cell lines, however, showed that one of the conserved elements – dubbed “island 2” – functioned as an allele-specific repressor: the protective allele A (but not the risk allele G) of SNP rs16888589 repressed luciferase expression below the level seen with the promoter-only reporter construct.

Working on the assumption that the rs16888589-containing repressor element targets the nearest gene *EIF3H*, Pittman et al conducted experiments aimed at elucidating the effect of differential *EIF3H* expression in colorectal cancer cell lines. They found that knocking down gene expression reduced cell proliferation and colony formation in a soft agar assay, and that overexpressing *EIF3H* increased cell proliferation. This suggests the possible role of a colorectal cancer oncogene for *EIF3H*. To further support its relevance to the functional *cis*-regulatory variant rs16888589, 3C experiments demonstrated that the island 2 repressor physically interacts with the *EIF3H* promoter in colorectal cancer cell lines (Pittman et al. 2010). Taken together, these data imply that the risk G allele of rs16888589 destroys the functionality of its long-range *EIF3H* repressor element, likely increasing *EIF3H* expression and possibly influencing colorectal cancer risk.

5.4 Misuse of enhancer elements at translocation breakpoints

Translocations are large-scale mutations where two nonhomologous chromosomes become joined. Genomic instability – a characteristic of many tumors – results in an increased number of translocations, some of which can have oncogenic effects on cells. These recurrent abnormal karyotypes were among the first genetic alterations to be identified in cancer cells, as they were visible using classic cytogenic approaches. As technology progressed, it became clear that the specific chromosomal breakpoints of a translocation were key to determining its potential impact of cell growth and differentiation. Some oncogenic translocations join the coding sequence of two different genes, generating fusion protein capable of promoting tumorigenesis. Others result from the juxtaposition of one gene's regulatory landscape (long-range *cis*-regulatory element[s]) with the coding sequence of another gene. Enhancers are promiscuous elements, capable of interacting with any promoter that enters their range of influence. This promiscuity allows for the improper activation of a gene outside its normal spatial range; this second example falls within the bounds of *cis*-regulatory variation underlying cancer etiology, as it involves the change to a gene's expression pattern due to alterations in its regulatory control.

5.4.1 Immunoglobulin translocations and heamatologic cancers

Recurrent translocations between the immunoglobulin (Ig) loci and assorted oncogene-partners is a hallmark of many leukemia and lymphoma cancers and a seminal example of aberrant oncogene transactivation due to chromosomal translocation (Nambiar et al. 2008; Willis and Dyer 2000). During normal B-cell development, the Ig heavy- and light-chain genes (IgH and IgL) undergo a process of rearrangement to produce a functional surface antigen receptor. These rearrangements are mediated by carefully controlled double-stranded DNA breaks (Kuppers 2005; Willis and Dyer 2000). While the mechanisms vary between cancer types and in many cases the precise pathogenesis of Ig translocations remain unclear, it is thought that many of the oncogenic translocations occur as mistakes during V(D)J recombination or during class-switching recombination (Kuppers 2005). Regardless of their mechanistic origins, these recurrent chromosomal rearrangements result in the juxtaposition of the active Ig *cis*-regulatory landscape and the coding portion of a given proto-oncogene, causing the production of a deregulated constitutively active oncogene in B-cells.

The t(14;18)(q32;q21) translocation is the most common chromosomal rearrangement in low-grade lymphomas (Duan et al. 2008). Its consequence is to bring the anti-apoptotic proto-oncogene *bcl-2* from chromosome 18q21 to the IgH locus on 14q32, yielding a deregulated and overexpressed *bcl-2* gene. Prolonged cell survival due to this misexpression has been shown to contribute to the development of lymphomas (Desoize 1994). While this common translocation was originally identified using cytogenetic approaches decades ago, work performed during the last several years has been crucial to uncovering the *cis*-regulatory elements and mechanisms through which the IgH regulatory landscape influences *bcl-2* misexpression.

The IgH locus harbors a cluster of long-range enhancer elements (the 3' IgH enhancers) comprised of four DNase I hypersensitive sites; these elements have been shown to function as a locus control region in B cells (Khamlichi et al. 2000). Direct evidence for the 3' IgH enhancers involvement in misregulating *bcl-2* first came from reporter gene assays in cell lines linking the 3' IgH enhancers directly to the *bcl-2* promoter; these constructs recapitulated the deregulation observed in lymphomas, with the Ig *cis*-elements driving high levels of expression and mimicking a *bcl-2* promoter usage shift seen *in vivo* (Duan et al. 2007). The enhancer elements are 350kb away from the translocation breakpoint *in vivo*, however, and the question of how they mediated *bcl-2* expression remained.

With the advent of 3C technology, Duan et al asked whether the 3' IgH enhancers were capable of looping to physically interact with the *bcl-2* promoter in t(14;18)(q32;q21) cells (Duan et al. 2008). Using two lymphoma cell lines – one with the translocation and one without – the authors looked for interactions between probes at the *bcl-2* promoter and those located in and around the 3' IgH enhancer cluster. They found that the two loci do indeed physically interact in the lymphoma line harboring the translocation, and that the interaction signal dropped of quickly outside of the enhancer cluster. Furthermore, they demonstrated that treatment with a drug known to decrease *bcl-2* transcription from the translocated locus (trichostatin A) dramatically decreased the IgH enhancer/ *bcl-2* promoter interaction as measured by 3C (Duan et al. 2005; Duan et al. 2008). This correlation between 3'IgH enhancer looping and *bcl-2* expression provides strong evidence for the enhancers' direct role in modulating *bcl-2* deregulation.

The gold standard for any functional hypothesis is to create a mouse model that recapitulates the desired phenotype. Xiang et al were able to do just that by showing that the introduction of the 3' IgH enhancers into the endogenous mouse *bcl2* locus caused *bcl-2* deregulation and the formation of follicular lymphomas (Xiang et al. 2011). Using mouse embryonic stem (ES) cells, they knocked the sequence surrounding the 3' IgH enhancers into the 3' region of the *bc-l2*, approximately 170kb downstream of the *bcl-2* promoter. The authors then characterized the mice, demonstrating an increase in B cell-specific *bcl-2* overexpression, extended B cell survival, and a physical interaction between the endogenous *bcl-2* promoter and the knocked-in 3' IgH enhancers. Finally, they showed that the mice developed B cell lymphomas (Xiang et al. 2011). These results conclusively prove that the 3' IgH enhancers are the *cis*-regulatory elements functionally responsible for the misregulation of *bcl-2* seen in t(14;18)(q32;q21) translocation.

5.4.2. TMPRSS2/ETS transcription factor translocations and prostate cancer

The oncogenic misexpression of proteins due to translocation is a signature of heamatologic cancers, and very few recurrent chromosomal arrangements have been identified in solid tumors (Mitelman 2000). One exception is a translocation commonly seen in prostate cancers that juxtaposes the 5' untranslated region of the chromosome 21q22.2 gene *TMPRSS2* – and all of the *cis*-regulatory elements contained within – with members of the *ETS* transcription factor gene family (Kumar-Sinha et al. 2008). *ETS* transcription factors are key proto-oncogenes involved in the control of cell growth, cell cycle regulation, and apoptosis, and are known to be overexpressed in numerous cancers (Hsu et al. 2004). Tomlins et al first identified this translocation by searching for “outlier” genes characterized by relatively low expression in most prostate cancer microarray profiles but highly overexpressed in a small percent of samples (Tomlins et al. 2005). Two *ETS* family transcription factors, v-ETS erythroblastosis virus E26 oncogene homolog (*ERG*) and ETS variant 1 (*ETV1*), appeared in their analysis. The authors investigated the nature of the *ERG* and *ETV1* overexpression in prostate cancer cell lines and specimens by performing exon-walking qPCR, where the expression level of each exon was interrogated individually. They noted that for both genes, the 5’ exon(s) were expressed at a reduced level compared to the rest of the protein; this suggested the presence of a translocation breakpoint between the normally expressed exon(s) and the downstream overexpressed neighbors. By using 5’ RNA ligase-mediated rapid amplification of cDNA ends (RACE) technology, they were able to discover that the 5’ exon(s) of *ERG* and *ETV1* had been replaced with the 5’ untranslated region of *TMPRSS2* (Tomlins et al. 2005). These two translocations were confirmed using fluorescence *in situ* hybridization (FISH), a technique that allows for the visualization of marked chromosomal locations in interphase cell spreads.

Transmembrane protease, serine 2 (*TMPRSS2*) is a prostate-specific, androgen-responsive gene that is expressed in both normal and neoplasic prostate tissue (Lin et al. 1999). The *ETS* gene translocations result in a fused transcript consisting of the 5’ untranslated first exon of *TMPRSS2* and the *ERG* or *ETV1* gene body; so while this translocation technically creates gene fusion products, there is no actual coding contribution from *TMPRSS2* (Kumar-Sinha et al. 2008). Instead, it is the *TMPRSS2* promoter and other *cis*-regulatory elements contained within the 5’ untranslated region and further upstream that cause the misexpression of the *ERG* or *ETV1* transcripts.

Work in cell lines and transgenic mice suggests that the *ETS* gene overexpression may result in increased invasiveness, suggesting a mechanism through which the translocation could mechanistically influence prostate cancer progression (Kumar-Sinha et al. 2008). *ERG* is the most commonly overexpressed oncogene in prostate cancer (Petrovics et al. 2005), and the *TMPRSS2* translocation was found to be present in 90% of cases exhibiting overexpression of *ERG* or *ETV1* (Tomlins et al. 2005). Therefore, this *cis*-regulatory gene fusion may underlie *ETS* oncogenic overexpression in the majority of prostate cancer cases.

5.5. Summary

Cancer, a disease of uncontrolled cellular proliferation, occurs when the genes normally responsible for regulating cell growth and division become misexpressed and cells gain the ability to bypass crucial cell cycle checkpoints. This overexpression of growth-promoting proto-oncogenes or underexpression of growth-curbing tumor suppressor genes can be caused by a plethora of different genetic mechanisms, and often the same key genes are subject to a variety of independent alterations. One means of tumorigenic misexpression is through mutations or variations affecting *cis*-regulatory elements. As described here, such *cis*-regulatory changes are involved in the etiology of many different cancers, and may help to explain the genetic underpinnings of these complex diseases. Recently, GWAS have been instrumental in identifying common risk variants in non-coding regions; functional follow-ups to these associations have resulted in the characterization of alternations in many *cis*-regulatory elements affecting the expression of nearby tumorigenic genes. Whether in the promoter, long-range elements such as enhancers or silencers, or in the overall architecture of a gene's *cis*-regulatory landscape, these mutations and variants have taught us much about the role of non-coding changes to cancer risk and progression.

While these *cis*-regulatory changes can have profound affects on gene expression, they are only one component of tumorigenic gene misexpression. Previously touched upon were other mechanisms that alter DNA sequence or structure: mutations to coding sequence, large-scale deletions or duplications, or translocations that create fusion proteins. Another facet of gene regulation – namely epigenetic marks and their dynamics – will also prove critical to understanding cancer etiology. While this type of variation has no impact on DNA sequence, it is likely to be at least as crucial as variation in non-coding DNA as a causative agent in tumorigenesis, and may help provide a link between the environmental factors known to play a role in cancer risk and actual gene expression changes. It is already well understood that cancer cells have profound methylation changes at many promoters (Esteller 2008; Feinberg and Tycko 2004), and the chromatin marks that help to define active and closed *cis*-regulatory elements and domains will also likely be linked to oncogenic misexpression. Future research will likely uncover the mechanisms linking epigenetics and cancer, enriching our understanding of the full impact *cis*-regulatory alterations have on tumorigenesis.

Ahmadiyeh N, Pomerantz MM, Grisanzio C, Herman P, Jia L, Almendro V, He HH, Brown M, Liu XS, Davis M et al. 2010. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc Natl Acad Sci U S A* **107:** 9742-9746.

Al Olama AA, Kote-Jarai Z, Giles GG, Guy M, Morrison J, Severi G, Leongamornlert DA, Tymrakiewicz M, Jhavar S, Saunders E et al. 2009. Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet* **41:** 1058-1060.

Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, Agnarsson BA, Sigurdsson A, Benediktsdottir KR, Cazier JB, Sainz J et al. 2006. A common variant associated with prostate cancer in European and African populations. *Nat Genet* **38:** 652-658.

Beke L, Nuytten M, Van Eynde A, Beullens M, and Bollen M. 2007. The gene encoding the prostatic tumor suppressor PSP94 is a target for repression by the Polycomb group protein EZH2. *Oncogene* **26:** 4590-4595.

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M et al. 2010. The landscape of somatic copy-number alteration across human cancers. *Nature* **463:** 899-905.

Bienz M and Clevers H. 2000. Linking colorectal cancer to Wnt signaling. *Cell* **103:** 311-320.

Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, Rowan A, Lubbe S, Spain S, Sullivan K, Fielding S et al. 2007. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet* **39:** 1315-1317.

Buckland PR, Hoogendoorn B, Coleman SL, Guy CA, Smith SK, and O'Donovan MC. 2005. Strong bias in the location of functional promoter polymorphisms. *Hum Mutat* **26:** 214-223.

Chang BL, Cramer SD, Wiklund F, Isaacs SD, Stevens VL, Sun J, Smith S, Pruett K, Romero LM, Wiley KE et al. 2009. Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk. *Hum Mol Genet* **18:** 1368-1375.

Chen Y and Olopade OI. 2008. MYC in breast tumor progression. *Expert Rev Anticancer Ther* **8:** 1689-1698.

Crowther-Swanepoel D, Broderick P, Di Bernardo MC, Dobbins SE, Torres M, Mansouri M, Ruiz-Ponte C, Enjuanes A, Rosenquist R, Carracedo A et al. 2010. Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat Genet* **42:** 132-136.

Curtin K, Lin WY, George R, Katory M, Shorto J, Cannon-Albright LA, Bishop DT, Cox A, and Camp NJ. 2009. Meta association of colorectal cancer confirms risk alleles at 8q24 and 18q21. *Cancer Epidemiol Biomarkers Prev* **18:** 616-621.

Dekker J, Rippe K, Dekker M, and Kleckner N. 2002. Capturing chromosome conformation. *Science* **295:** 1306-1311.

DeMarzo AM, Nelson WG, Isaacs WB, and Epstein JI. 2003. Pathological and molecular aspects of prostate cancer. *Lancet* **361:** 955-964.

Desoize B. 1994. Anticancer drug resistance and inhibition of apoptosis. *Anticancer Res* **14:** 2291-2294.

Duan H, Heckman CA, and Boxer LM. 2005. Histone deacetylase inhibitors down-regulate bcl-2 expression and induce apoptosis in t(14;18) lymphomas. *Mol Cell Biol* **25:** 1608-1619.

Duan H, Heckman CA, and Boxer LM. 2007. The immunoglobulin heavy-chain gene 3' enhancers deregulate bcl-2 promoter usage in t(14;18) lymphoma cells. *Oncogene* **26:** 2635-2641.

Duan H, Xiang H, Ma L, and Boxer LM. 2008. Functional long-range interactions of the IgH 3' enhancers with the bcl-2 promoter region in t(14;18) lymphoma cells. *Oncogene* **27:** 6720-6728.

Easton DF Pooley KA Dunning AM Pharoah PD Thompson D Ballinger DG Struewing JP Morrison J Field H Luben R et al. 2007. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447:** 1087-1093.

Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, Mulholland S, Leongamornlert DA, Edwards SM, Morrison J et al. 2008. Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* **40:** 316-321.

Esteller M. 2008. Epigenetics in cancer. *N Engl J Med* **358:** 1148-1159.

Feinberg AP and Tycko B. 2004. The history of cancer epigenetics. *Nat Rev Cancer* **4:** 143-153.

Ghoussaini M, Song H, Koessler T, Al Olama AA, Kote-Jarai Z, Driver KE, Pooley KA, Ramus SJ, Kjaer SK, Hogdall E et al. 2008. Multiple loci with different cancer specificities within the 8q24 gene desert. *J Natl Cancer Inst* **100:** 962-966.

Grigoriadis A, Mackay A, Reis-Filho JS, Steele D, Iseli C, Stevenson BJ, Jongeneel CV, Valgeirsson H, Fenwick K, Iravani M et al. 2006. Establishment of the epithelial-specific transcriptome of normal and malignant human breast cells based on MPSS and array expression data. *Breast Cancer Res* **8:** R56.

Grisanzio C and Freedman ML. 2010. Chromosome 8q24-Associated Cancers and MYC. *Genes & Cancer* **1:** 555-559.

Gudmundsson J, Sulem P, Gudbjartsson DF, Jonasson JG, Sigurdsson A, Bergthorsson JT, He H, Blondal T, Geller F, Jakobsdottir M et al. 2009. Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. *Nat Genet* **41:** 460-464.

Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A et al. 2007. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* **39:** 631-637.

Haiman CA, Le Marchand L, Yamamato J, Stram DO, Sheng X, Kolonel LN, Wu AH, Reich D, and Henderson BE. 2007a. A common genetic risk factor for colorectal and prostate cancer. *Nat Genet* **39:** 954-956.

Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, Waliszewska A, Neubauer J, Tandon A, Schirmer C, McDonald GJ et al. 2007b. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet* **39:** 638-644.

He TC, Sparks AB, Rago C, Hermeking H, Zawel L, da Costa LT, Morin PJ, Vogelstein B, and Kinzler KW. 1998. Identification of c-MYC as a target of the APC pathway. *Science* **281:** 1509-1512.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39:** 311-318.

Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S, Chandler I, Vijayakrishnan J, Sullivan K, Penegar S et al. 2008. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* **40:** 1426-1435.

Hsu T, Trojanowska M, and Watson DK. 2004. Ets proteins in biological control and cancer. *J Cell Biochem* **91:** 896-903.

Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A et al. 2007. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39:** 870-874.

Jia L, Landan G, Pomerantz M, Jaschek R, Herman P, Reich D, Yan C, Khalid O, Kantoff P, Oh W et al. 2009. Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet* **5:** e1000597.

Katoh M. 2008. Cancer genomics and genetics of FGFR2 (Review). *Int J Oncol* **33:** 233-237.

Khamlichi AA, Pinaud E, Decourt C, Chauveau C, and Cogne M. 2000. The 3' IgH regulatory region: a complex structure in a search for a function. *Adv Immunol* **75:** 317-345.

Kiemeney LA, Thorlacius S, Sulem P, Geller F, Aben KK, Stacey SN, Gudmundsson J, Jakobsdottir M, Bergthorsson JT, Sigurdsson A et al. 2008. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat Genet* **40:** 1307-1312.

Kumar-Sinha C, Tomlins SA, and Chinnaiyan AM. 2008. Recurrent gene fusions in prostate cancer. *Nat Rev Cancer* **8:** 497-511.

Kuppers R. 2005. Mechanisms of B-cell lymphoma pathogenesis. *Nat Rev Cancer* **5:** 251-262.

Landa I, Ruiz-Llorente S, Montero-Conde C, Inglada-Perez L, Schiavi F, Leskela S, Pita G, Milne R, Maravall J, Ramos I et al. 2009. The variant rs1867277 in FOXE1 gene confers thyroid cancer susceptibility through the recruitment of USF1/USF2 transcription factors. *PLoS Genet* **5:** e1000637.

Levy L and Hill CS. 2006. Alterations in components of the TGF-beta superfamily signaling pathways in human cancer. *Cytokine Growth Factor Rev* **17:** 41-58.

Lin B, Ferguson C, White JT, Wang S, Vessella R, True LD, Hood L, and Nelson PS. 1999. Prostate-localized and androgen-regulated expression of the membrane-bound serine protease TMPRSS2. *Cancer Res* **59:** 4180-4184.

Lou H, Yeager M, Li H, Bosquet JG, Hayes RB, Orr N, Yu K, Hutchinson A, Jacobs KB, Kraft P et al. 2009. Fine mapping and functional analysis of a common variant in MSMB on chromosome 10q11.2 associated with prostate cancer susceptibility. *Proc Natl Acad Sci U S A* **106:** 7933-7938.

Meyer KB, Maia AT, O'Reilly M, Teschendorff AE, Chin SF, Caldas C, and Ponder BA. 2008. Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS Biol* **6:** e108.

Middeldorp A, Jagmohan-Changur S, van Eijk R, Tops C, Devilee P, Vasen HF, Hes FJ, Houlston R, Tomlinson I, Houwing-Duistermaat JJ et al. 2009. Enrichment of low penetrance susceptibility loci in a Dutch familial colorectal cancer cohort. *Cancer Epidemiol Biomarkers Prev* **18:** 3062-3067.

Mitelman F. 2000. Recurrent chromosome aberrations in cancer. *Mutat Res* **462:** 247-253.

Nambiar M, Kari V, and Raghavan SC. 2008. Chromosomal translocations in cancer. *Biochim Biophys Acta* **1786:** 139-152.

Nesbit CE, Tersak JM, and Prochownik EV. 1999. MYC oncogenes and human neoplastic disease. *Oncogene* **18:** 3004-3016.

Nobrega MA, Ovcharenko I, Afzal V, and Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302:** 413.

Onodera Y, Miki Y, Suzuki T, Takagi K, Akahira J, Sakyu T, Watanabe M, Inoue S, Ishida T, Ohuchi N et al. 1111. Runx2 in human breast carcinoma: its potential roles in cancer progression. *Cancer Sci* **101:** 2670-2675.

Parlato R, Rosica A, Rodriguez-Mallon A, Affuso A, Postiglione MP, Arra C, Mansouri A, Kimura S, Di Lauro R, and De Felice M. 2004. An integrated regulatory network controlling survival and migration in thyroid organogenesis. *Dev Biol* **276:** 464-475.

Petrovics G, Liu A, Shaheduzzaman S, Furusato B, Sun C, Chen Y, Nau M, Ravindranath L, Dobi A, Srikantan V et al. 2005. Frequent overexpression of ETS-related gene-1 (ERG1) in prostate cancer transcriptome. *Oncogene* **24:** 3847-3852.

Pittman AM, Naranjo S, Jalava SE, Twiss P, Ma Y, Olver B, Lloyd A, Vijayakrishnan J, Qureshi M, Broderick P et al. 2010. Allelic variation at the 8q23.3 colorectal cancer risk locus functions as a cis-acting regulator of EIF3H. *PLoS Genet* **6**.

Pittman AM, Naranjo S, Webb E, Broderick P, Lips EH, van Wezel T, Morreau H, Sullivan K, Fielding S, Twiss P et al. 2009. The colorectal cancer risk at 18q21 is caused by a novel variant altering SMAD7 expression. *Genome Res* **19:** 987-993.

Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA, Chan JA, Hills A, Davis M et al. 2009. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41:** 882-884.

Pomerantz MM, Shrestha Y, Flavin RJ, Regan MM, Penney KL, Mucci LA, Stampfer MJ, Hunter DJ, Chanock SJ, Schafer EJ et al. 2010. Analysis of the 10q11 cancer risk locus implicates MSMB and NCOA4 in human prostate tumorigenesis. *PLoS Genet* **6:** e1001204.

Reeves JR, Dulude H, Panchal C, Daigneault L, and Ramnani DM. 2006. Prognostic value of prostate secretory protein of 94 amino acids and its binding protein after radical prostatectomy. *Clin Cancer Res* **12:** 6018-6022.

Rivas M, Mellstrom B, Torres B, Cali G, Ferrara AM, Terracciano D, Zannini M, Morreale de Escobar G, and Naranjo JR. 2009. The DREAM protein is associated with thyroid enlargement and nodular development. *Mol Endocrinol* **23:** 862-870.

Sequeira MJ, Morgan JM, Fuhrer D, Wheeler MH, Jasani B, and Ludgate M. 2001. Thyroid transcription factor-2 gene expression in benign and malignant thyroid lesions. *Thyroid* **11:** 995-1001.

Sotelo J, Esposito D, Duhagon MA, Banfield K, Mehalko J, Liao H, Stephens RM, Harris TJ, Munroe DJ, and Wu X. 2010. Long-range enhancers on 8q24 regulate c-Myc. *Proc Natl Acad Sci U S A* **107:** 3001-3005.

ten Dijke P and Hill CS. 2004. New insights into TGF-beta-Smad signalling. *Trends Biochem Sci* **29:** 265-273.

Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, Barnetson RA, Theodoratou E, Cetnarskyj R, Cartwright N et al. 2008. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* **40:** 631-637.

Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, Yu K, Chatterjee N, Welch R, Hutchinson A et al. 2008. Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet* **40:** 310-315.

Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R et al. 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310:** 644-648.

Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M, Wood W et al. 2007. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* **39:** 984-988.

Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, Spain S, Lubbe S, Walther A, Sullivan K et al. 2008. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* **40:** 623-630.

Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS et al. 2010. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet* **42:** 504-507.

Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Bjorklund M, Wei G, Yan J, Niittymaki I et al. 2009. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* **41:** 885-890.

Udler MS, Meyer KB, Pooley KA, Karlins E, Struewing JP, Zhang J, Doody DR, MacArthur S, Tyrer J, Pharoah PD et al. 2009. FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Hum Mol Genet* **18:** 1692-1703.

Visel A, Rubin EM, and Pennacchio LA. 2009. Genomic views of distant-acting enhancers. *Nature* **461:** 199-205.

Wasserman NF, Aneas I, and Nobrega MA. 2010. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res* **20:** 1191-1197.

Willis TG and Dyer MJ. 2000. The role of immunoglobulin translocations in the pathogenesis of B-cell malignancies. *Blood* **96:** 808-822.

Wright JB, Brown SJ, and Cole MD. 2010. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol Cell Biol* **30:** 1411-1420.

Xiang H, Noonan EJ, Wang J, Duan H, Ma L, Michie S, and Boxer LM. 2011. The immunoglobulin heavy chain gene 3' enhancers induce Bcl2 deregulation and lymphomagenesis in murine B cells. *Leukemia* **2011:** 24.

Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N et al. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39:** 645-649.

Yeager M, Xiao N, Hayes RB, Bouffard P, Desany B, Burdett L, Orr N, Matthews C, Qi L, Crenshaw A et al. 2008. Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. *Hum Genet* **124:** 161-170.

Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E et al. 2007. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* **39:** 989-994.

Figure 1. Strategies to map genetic variation affecting disease traits due to changes in gene expression in human populations. A. Genome Wide Association Studies (GWAS) identify genetic variants (SNPs) associated with a disease trait. Differently than most SNPs in the genome, which have similar allele frequencies (red and green individuals) in affected (cases) and non-affected (controls) individuals, an associated variant shows a significant departure from this pattern; in the example shown, there is an overabundance of the “red” allele of the associated SNP in cases, compared to “green” alleles in controls. **B.** The associated variant in not necessarily the causal variant underlying the phenotypic difference; rather, multiple SNPs are highly correlated with one another in Linkage Disequilibrium Blocks (LD blocks). Various strategies are used to identify which SNPs (red asterisk) within these LD blocks might have a putatively causal role in the phenotype-genotype association. For example, SNPs mapping within evolutionarily conserved noncoding sequences (green peaks along the LD block) are good candidates for having a role in phenotypic variation. Further analysis of the genomic context of this candidate SNP can further support the idea that this variant lies within a cis-regulatory element, showing, for example, that the local chromatin is compatible with that seen in active cis-regulatory elements (single green balls on the histones, denoted as blue balls). For genome-wide chromatin states in multiple cell lines, see the ENCODE project data at http://genome.ucsc.edu/cgi-bin/hgGateway. More detailed computational analysis may reveal that the SNP lies within a well defined DNA binding motif for a given transcription factor. This raises the hypothesis that the SNP may alter the binding of proteins to a cis-regulatory element, resulting in differential gene expression. **C.** Multiple experimental strategies can be used to determine that a cis-regulatory element controls the expression of a given gene and that a SNP within this regulatory sequence may alter its function. Electro Mobility Shift Assays (EMSA) are

used to show that a specific protein has the ability to bind to the given stretch of DNA containing the SNP in question (lane 2 of the gel). Chromatin Immunoprecipitation (ChIP) detects the binding of a transcription factor to a specific DNA sequence. Reporter assays can be used to test whether a given DNA sequence is an enhancer, and whether SNP within this enhancer may result in allele-specific functions. These reporter assays can employ *in vitro* or *in vivo* experimental models. Chromatin conformation capture (3C) demonstrates long-range interactions in the genome. A putative enhancer (green) loops to activate a distant promoter (blue) of a gene (red arrow). This looping can be captured by cross-linking (gray balls) followed by PCR using primers (black arrows) for the enhancer and the promoter. PCR amplification using these primers demonstrates that the two distant sequences directly interact, as predicted to occur between enhancers and their distant promoters.

Figure 2. How mutations affect gene expression. A. Endogenous expression pattern of a gene. **B.** A promoter variant increases overall gene expression levels. **C.** The long-range enhancer model: three tissue-specific enhancers determine normal gene expression. **D.** An inactivating mutation in a brain enhancer (yellow) results in a reduced expression domain. **E.** An activating variant in a second brain enhancer (orange) results in brain-specific overexpression. **F.** A translocation juxtaposes a limb enhancer (green) into the gene’s regulatory landscape, resulting in an expanded expression domain.

Figure 1.

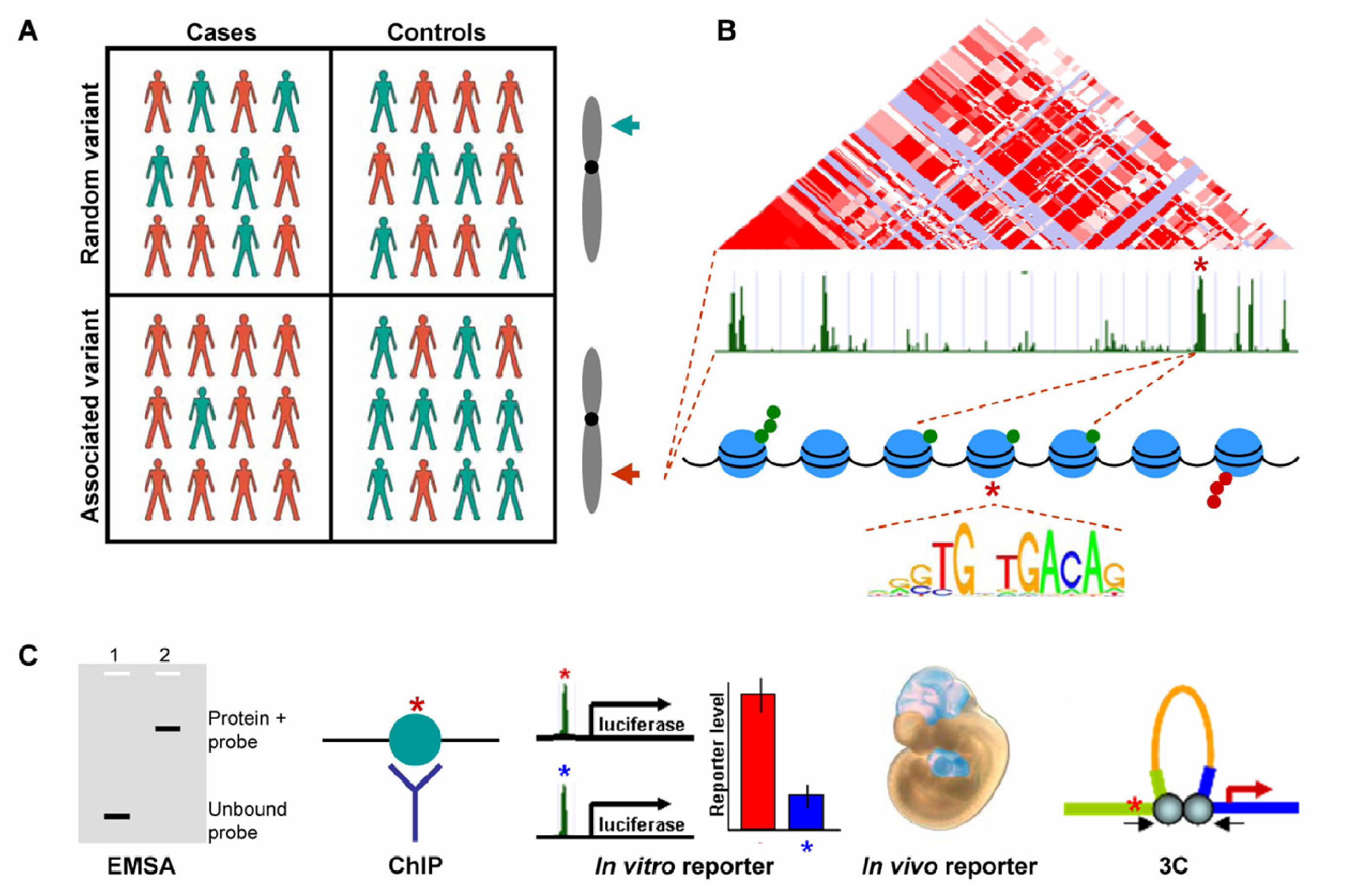
A
Cases
Controls
Random variant
Associated variant
B
C
1
2
Protein +
probe
Unbound
probe
EMSA
ChIP
luciferase
luciferase
Reporter level
In vitro reporter
In vivo reporter
3C

Figure 2.